Spring 2014

Opinion Mining on Twitter Data Stream to Give Companies an Up-to-Date Feedback on Their Free Products

Lokmanyathilak Govindan Sankar Selvan
San Jose State University

Opinion Mining on Twitter Data Stream to Give Companies an Up-to-Date Feedback on Their Free Products

A Writing Project
Presented to
The Faculty of the Department of Computer Science
San José State University
In Partial Fulfillment of the
Requirements for the
Degree Master of Computer Science
By

Lokmanyathilak Govindan Sankar Selvan

Spring 2014

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

Opinion Mining on Twitter Data Stream to Give Companies an Up-to-Date Feedback on Their Free Products

by

Lokmanyathilak Govindan Sankar Selvan

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

_____

Dr. Teng Moh, Department of Computer Science

_____

Dr. Mark Stamp, Department of Computer Science

_____

Dr. Sami Khuri, Department of Computer Science

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

ABSTRACT

There are lots of companies producing various products ranging from expensive to free products. There is no software product without any bug irrespective of their cost. The problem with this situation is that when people purchase a software product by paying money they are more concerned about its performance. People report to the companies if the product they purchased does not work as expected. It is not the same in the case of free products. People tend to switch to some other free product produced by different company which does the same job. The notion of this project is to solve this problem. Under such a circumstance, people may not send a report to the company but people talk about that in the social networks. In this project, we are going to make use of this kind of posts or tweets to help the companies by informing them about the fault in their free products so that their reputation will never go down. So the objective of this project is to stream the real-time Twitter data, filter the data and analyze the data so that it can be reported to the companies if their free products does not work as expected.

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

ACKNOWLEDGEMENT

I would like to thank Dr. Teng Moh for his technical guidance and his constant support during the writing project. Dr. Moh's timely advice and his expertise in Social Web have helped me all through this project. I would also like to thank Dr. Mark Stamp and Dr. Sami Khuri for being my committee members and help me through this process.

TABLE OF CONTENTS

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

LIST OF FIGURES

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

1. PROJECT OVERVIEW

1.1. INTRODUCTION

After so many years of advancement in science and technology, there are lot of software products in the market that fail miserably as they do not satisfy the customers. There are strategic teams and testing teams within the company who are working to produce a quality product for the customers. The strategic team take surveys and create strategies based on the surveys. Even in the present days we can see online surveys in many sites which are usually not liked by the customers. People don't like these surveys because they need to take an effort in filling them and they need to spend their time on it.

Opinion mining [1] is the best way to solve this problem. It is fast growing topic and lot of organizations are conducting research in this project as they feel there is a great chance of improving the market of a product if they are aware of people's pulse. Opinion mining can be done in any source of textual data. Due to the recent advancement in online forums and social networking sites it has become much easier to collect data. Twitter data stream is one of the best places to study and analyze them.

The whole idea of this project is to analyze a particular data stream from Twitter micro-blogging site so that it helps companies to have up-to-date knowledge about their free products.

1.2. SCENARIO AND INFERENCE

Scenario 1: Google chrome

It is one of the most used free products of Google. Let us consider a person using Google chrome for browsing and his entire internet surfing. Google chrome can crash for several reasons

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

like long time usage with many numbers of tabs, on installing some plug-ins, etc [2]. If Google chrome crashes often on his computer, a normal person will not be able to find the solution. His immediate solution would be to switch browser.

Scenario 2: Microsoft Silverlight

Microsoft Silverlight is similar to Adobe Flash. It is used to run media applications in the web browser. It can be installed to the web browser as a plug-in. Now consider a situation in which Silverlight crashes often on the web browser and the web browser starts showing not responding on the title bar [3]. In this scenario people tend to switch to Adobe Flash as it is also free.

This scenario not only lets the reputation of Microsoft down, it also pulls Netflix's reputation down.

Inference:

Common people never spend time in complaining about free products to the companies producing them. They have a much easy solution, which is moving on to some other free software which performs the same task.

Similarly think of situation in which Microsoft Office crashes often and is not working. In this case, person who has purchased it will definitely report about this to the company.

People who are really frustrated because of the performance of free products tend to show their rage by talking about those products in social networks.

2. RELATED WORK:

There are several ways we can handle data to reach the goal of this project. Most of the sentiment analysis tools or algorithms are still under research. Till date, there is no algorithm that can provide hundred percent accurate results for sentiment analysis. There are several debates going on between various researchers in this topic to prove their solution is more perfect than the others. English is simple but tricky language as many words present in this language can be used in more than one sense (that is, they can be used in a positive, negative as well as neutral sense). Therefore the analysis part becomes very hard. In this project we use a sentimental dictionary using which we can provide nine different weights to words in tweets varying from -4 to +4.

The tool or platform that we need should be versatile enough to handle all the problems in the data that we stream from a social networking site [4] [5]. Apache Hadoop [6] is chosen to deal with the huge volume of data which I stream from Twitter. Hadoop is designed in such a way that it can format the unstructured data into structured pattern. It is the perfect tool currently in the market to perform analytics over huge data. Data streamed from Twitter is unstructured and it is in JSON format [7]. Each and every array of JSON might be different as they are from various sources and different people. They can have different number of attributes. Apache Hive [8] is a component built over Hadoop that has a serializer and deserializer called JSON SerDe [9] which reads unstructured data and writes columnar structured data. This makes the process of analysis easy.

A unique flow of process is created by carefully merging different operations like data streaming, storage, formatting, filtering, analysis and presentation into a complete application.

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

3. SENTIMENTAL WORD DICTIONARY:

Sentimental word dictionary [10] becomes essential as we are dealing with text analysis. The existing sentimental dictionary has words which are categorized into positive, negative or neutral polarity. This dictionary will produce results which are far from accurate.

In this project, to improve the accuracy of the result, the words are further classified into weak, strong and very_strong based on their strength. As each and every strength has possibility of all three polarities, the number of categories based on this classification becomes nine. The words are further classified into parts of speech as noun, verb, adjective, adverb and others. By classifying the words in this pattern we can provide different weightage to the categories.

By making all these changes in the sentimental word dictionary, we can make the result of sentiment analysis close to accurate. This method can provide even more accurate results as we train it with all possible kinds of test data.

4. TOOLS AND FRAMEWORK:

4.1. HADOOP FRAMEWORK:

Apache Hadoop must be installed on a cloud server or a local host. Other services like Flume, Hive and Oozie should be installed over it. Cloudera version of Hadoop provides the services that makes the work easy.

4.1.1. FLUME, HIVE & OOZIE:

Apache Flume is a service which is used to stream large amount of real-time data. Flume is used in our project to stream real-time data from Twitter and we store it the Hadoop Distributed File System (HDFS) [11].

Hive warehouse is used to store data. Data can be managed and retrieved by querying. HiveQL [18] is the querying language used in Hive which has syntax similar to SQL. Hive parses the query and creates an execution plan in the form of tree of operators. Map-Reduce [12] jobs are triggered with the tree of operators is read recursively.

Apache Oozie [13] is used to maintain the workflow of data in and around Hive and HDFS.

The image below shows all the services including Flume, Hive, Oozie and HDFS are installed in the host using Cloudera Manager [14]. The Cloudera version of Apache Flume, Apache Hive and Apache Oozie are easy to install and operate due to the convenient interface provided by the Cloudera Manager.

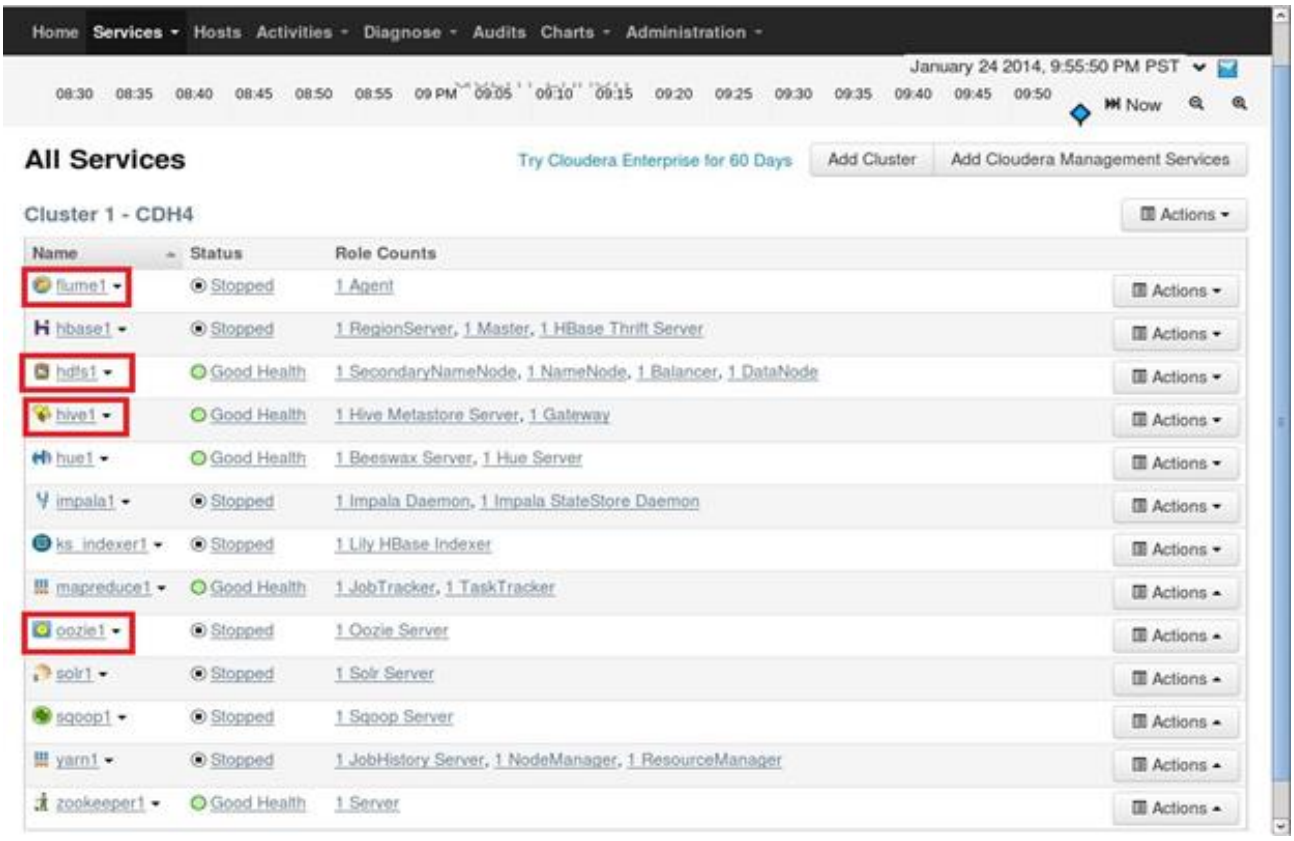


Figure 1: Cloudera Manager with all services installed

Twitter developer account is needed. A dummy app must be created and Twitter API [15] OAuth settings like key and secret can be used to retrieve Twitter data.

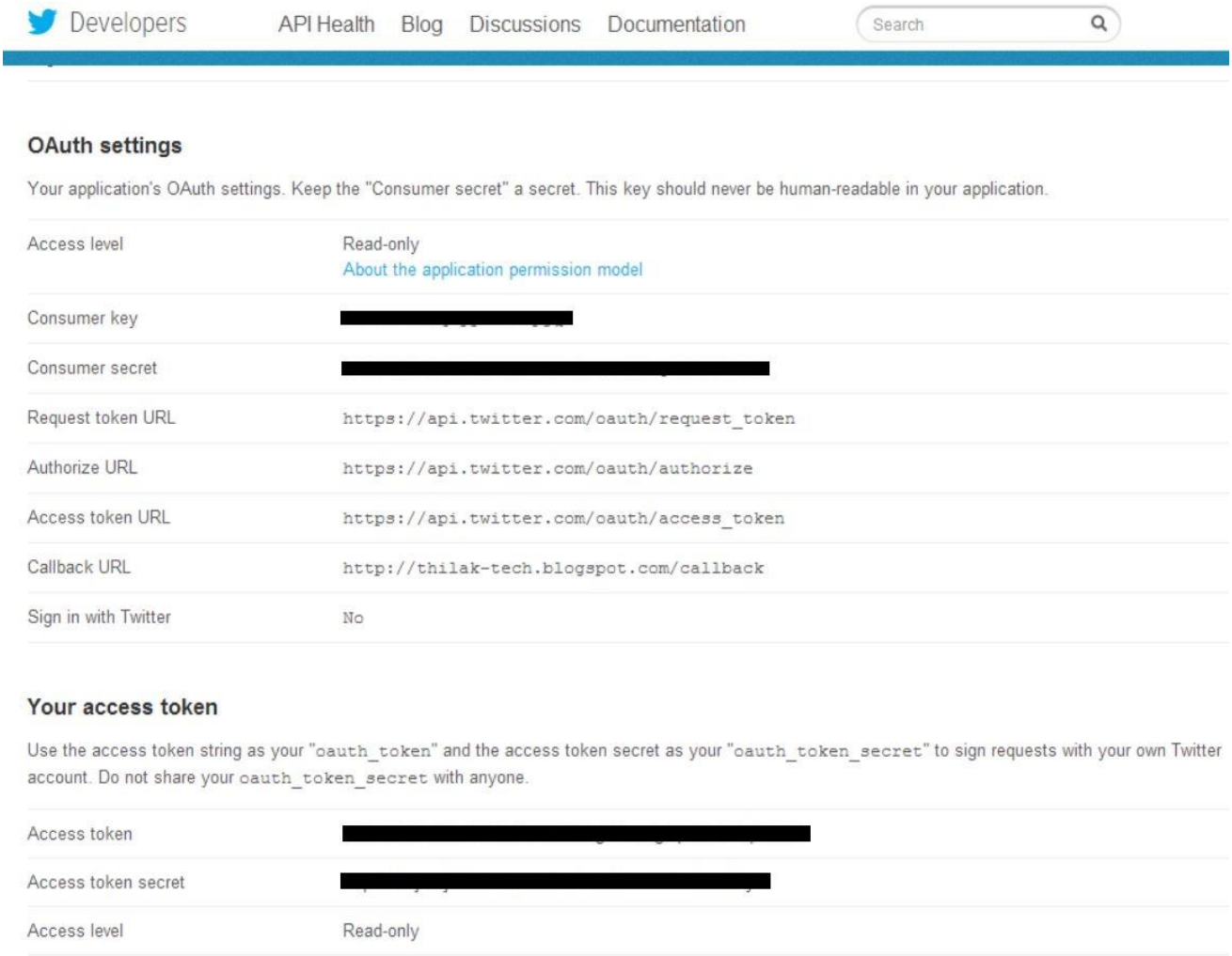


Figure 2: Twitter Developer Account – Twitter API Key and Secret

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

5. ARCHITECTURE

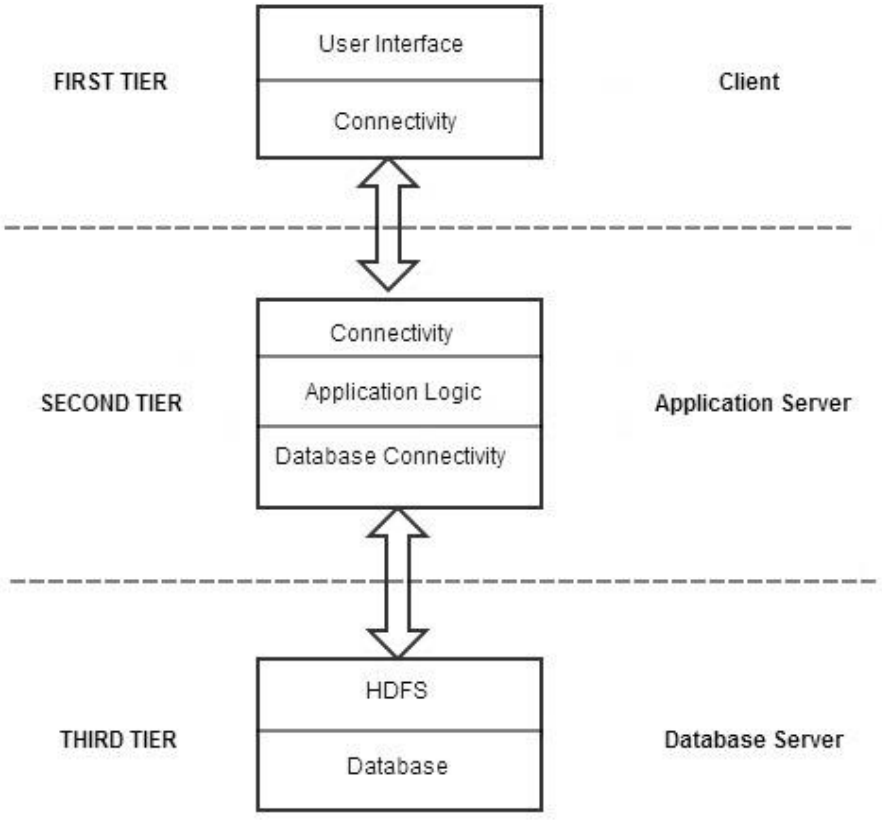


Figure 3: Architecture Diagram

This project is created in a three tier architecture. The first tier is the layer which is visible to the end-user. The input is given and the output is displayed in this layer. It is the client and it is connected to the next layer using connectivity drivers.

Second tier is the important layer which performs all the logical operations in this project. This is the middle layer and it acts as a medium between first and third layer. It is connected to the third layer to by the database connectivity.

The third tier is the storage layer. This layer has the HDFS which stores and manages the data which comes into the application.

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

6. SYSTEM DECOMPOSITION

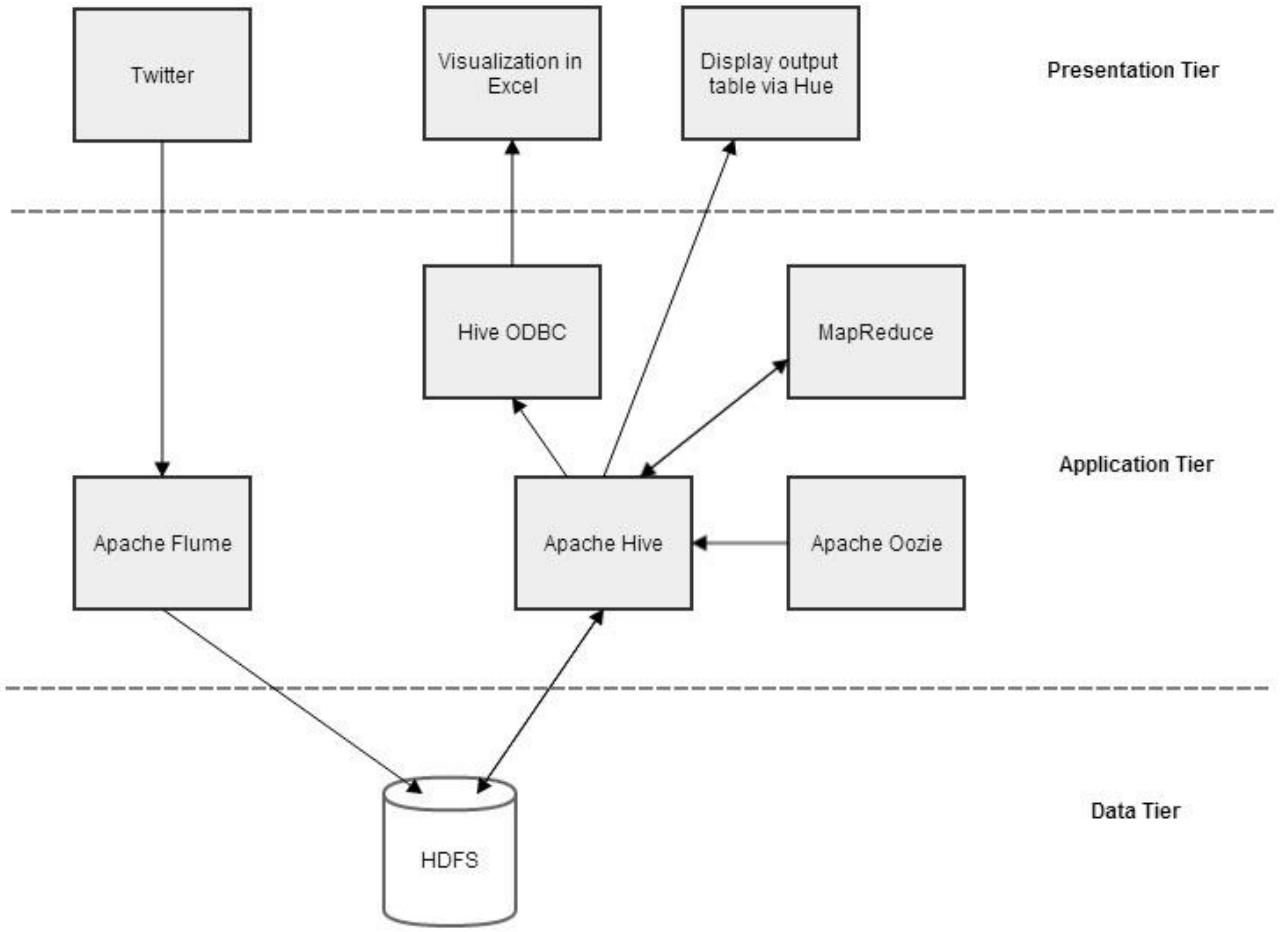


Figure 4: System Decomposition Diagram

Presentation Tier:

It is the topmost and interactive layer in this architecture. Input from Twitter is fed into the system from this layer. The input data is after getting processed is sent to the presentation layer. Output in the form of tables and graphs are displayed in this layer. The tables are viewed in via Hue interface which is a part of the Hadoop ecosystem.

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

Application Tier:

This layer lies between presentation and data tier. All the logical operations are performed here. Apache Flume, Hive and Oozie are the important components present in this layer. Flume component of Hadoop is the one which receives the tweets from Twitter data stream and sends it to the database. Apache Hive is the one which is very closely coupled to the HDFS. Hive metastore stores the information about all the data present in the database. Hive query language (Hive-QL) is used to process the data. All the process are performed by the map-reduce jobs and the map-reduce jobs are triggered by the Hive queries. Apache Oozie is used to create partitions in the tables which improves the performance of the querying process. Hive ODBC is used to export the results of the query into Windows operating system with which the visualizations are created in Microsoft Excel.

Data Tier:

This is the bottom most layer in this application. It receives data from Flume and stores in the Hadoop Distributed File System (HDFS). All the components present in the application tier needs to access the database to proceed with their process. Database connectivity should be established to transfer data. Hive establishes connections with the Hive metastore from where it begins all the transactions. HDFS can store large volume of data as it is a distributed file system. This type of file system is suitable only for very large data as the performance in low when compared to the traditional RDBMS.

7. PROJECT DESIGN

7.1. DESIGN OUTLINE

Storing and querying the tweets from Twitter in a traditional RDBMS is not possible because it is in JSON format and it is unstructured. Therefore it is too complex to be queried using RDBMS queries. The Hadoop ecosystem has the Hive project which deals with this problem. The Hive query language is similar to SQL but allows us to perform querying on complex data types.

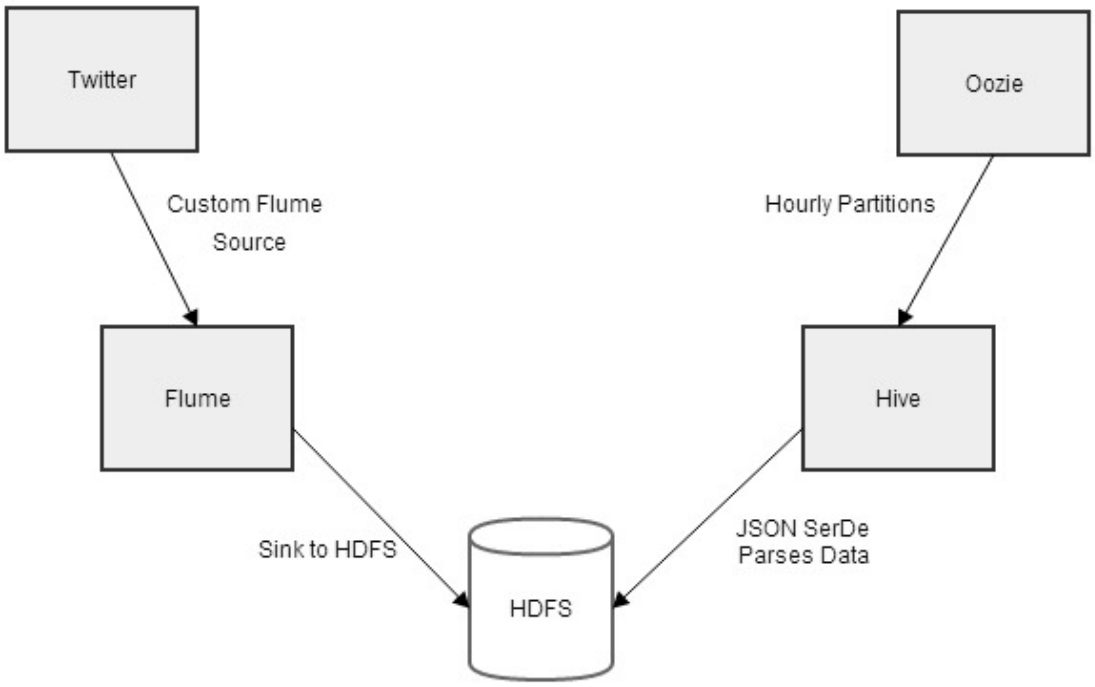


Figure 5: Data Flow

The first step is to collect data from Twitter through Twitter Streaming API [15] using Flume. Apache Flume [16] is configured in such a way that it know its endpoints which are sources and sinks. Each and every piece of tweet is called an event. The tweets are gathered from the Twitter

source and sent to the sink through the memory channel. The sinks are responsible for writing them to HDFS files.

The next step is to create and manage partitions with Oozie. An external table is created using the files in the HDFS. External tables are used so the querying can be done without moving the data from the location. Partitions are be created as the data gets into the database. This process can be automated as it improves the performance of the query.

The final step is querying complex data using Hive-QL [18].

The project can be divided into five steps, namely data streaming, data storage, data filtration, data analytics and finally reporting phase.

7.2. DATA STREAMING

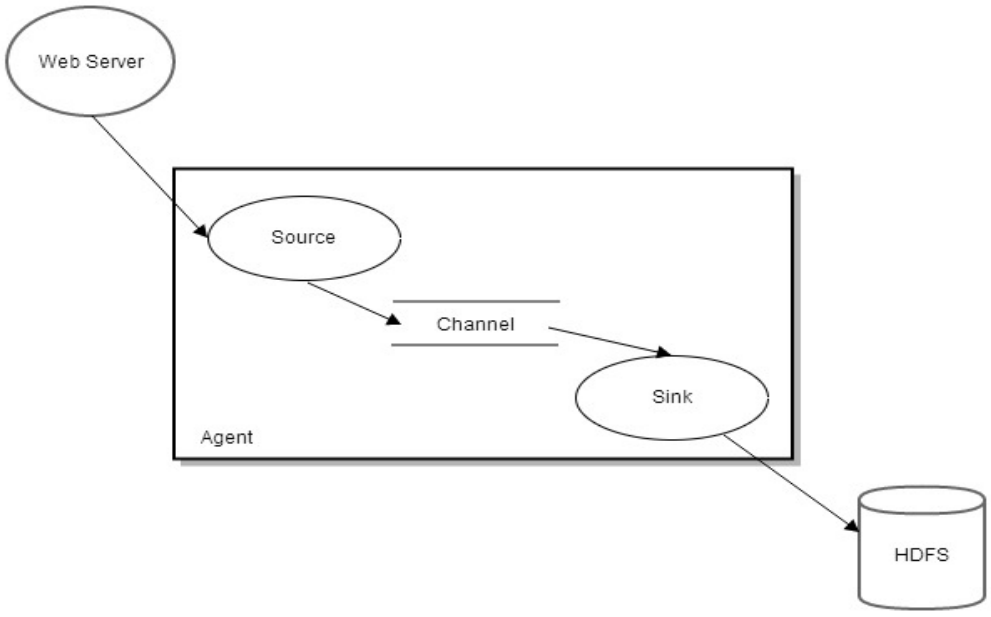


Figure 6: Data Streaming

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

Data streaming is performed using Flume, which is a component of the Hadoop ecosystem. Flume is used to stream real-time data. In the project we use Flume to stream real-time twitter data into Hadoop Distributed File System (HDFS) [11].

Flume collects the data from Twitter Streaming API [16] and forwards it to the HDFS. On taking a closer look at this process we get an image like the one below.

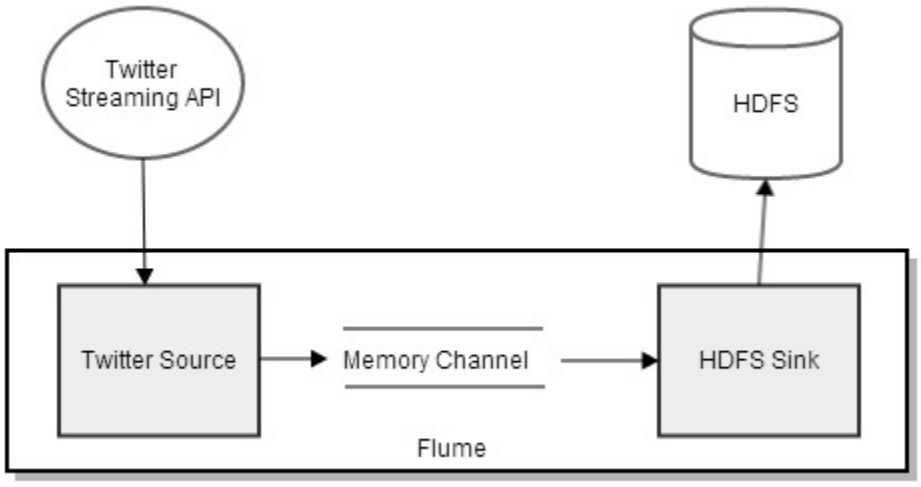


Figure 7: Flume

Twitter Streaming API:

HTTP server and streaming connection are the essential parts which make this happen. The user sends request to HTTP server. The streaming process connects to Twitter and receives the data. Then it parses the data and stores it as result. The HTTP server process sends the result from streaming connection process to the user as response. The connection is closed once the required data is received by the user.

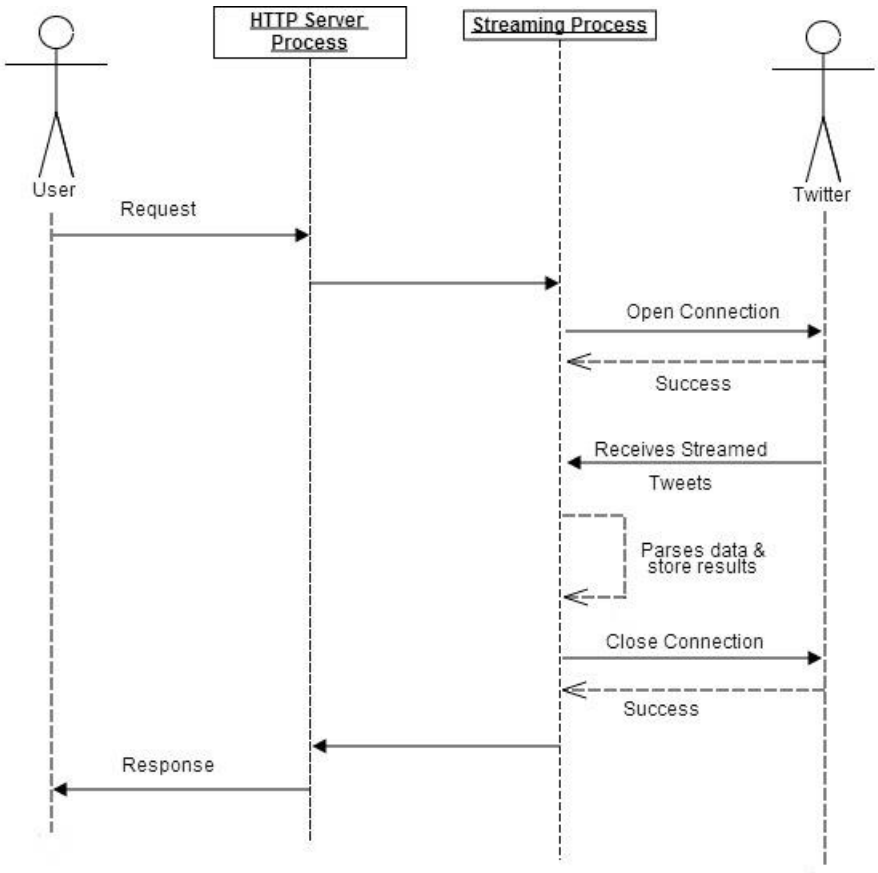


Figure 8: Twitter Streaming API

Source:

Source is used to get data from clients and the send them to the channel. In the project, the Twitter Source gets the data from Twitter Streaming API [16] which is the external client and stores it into the memory channel.

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

Channel:

Channel acts as the intermediate pathway between the Source and the Sink. Here, the memory channel transfers the data received from Twitter Source into HDFS Sink.

In this project, memory channel is defined in the flume configuration file.

```
TwitterAgent.channels.MemChannel.type = memory
```

Sink:

Sink extracts the Events from channel and then it sends the events to Flume agent. Flume configuration file is edited to provide the path where the streamed data is stored.

```
TwitterAgent.sinks.HDFS.hdfs.path =
hdfs://localdomain.local:8020/user/flume/tweets/%Y/%m/%d/%H/
```

8. PROJECT IMPLEMENTATION

There are five main stages in the implementation of this project. They are data streaming, data storage, data filtration, data analytics and data presentation. The overview of all these stages are shown in the image below.

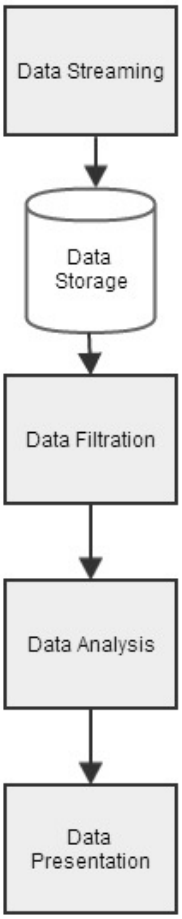


Figure 9: Overall Process

8.1. CONFIGURING HADOOP AND STREAMING DATA

Configuring Flume Agent:

The Flume configuration file looks like,

```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS
```


Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

```
TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = *******
TwitterAgent.sources.Twitter.consumerSecret = *******
TwitterAgent.sources.Twitter.accessToken = *******
TwitterAgent.sources.Twitter.accessTokenSecret = *******
TwitterAgent.sources.Twitter.keywords = Google chrome
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path =
hdfs://localdomain.local:8020/user/flume/tweets/%Y/%m/%d/%H/
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
TwitterAgent.sinks.HDFS.hdfs.rollInterval = 600

TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

The Flume Agent is configured so that the Twitter data can be streamed into the HDFS when the Twitter Agent is started.

Twitter API key and secret are given as '*******' in the above image as they should not be shared. Keyword is given as 'Google chrome' so that it streams all the tweets with the keyword in it. Many number of keywords can be added if necessary. The streamed data is stored in the HDFS in the directory created by the path in the code above. Year, month, day and hour directory will be created automatically within the HDFS as the data is streamed into it. The directories are created in such way to avoid confusion and increase performance.

Starting the Twitter Agent:

The Flume starts streaming the real-time Twitter data after starting the Twitter Agent. The Twitter Agent is started by the command below.

```
$ /etc/init.d/flume-ng-agent start
```

Configuring MySQL database for Hive Metastore:

Before proceeding with the configuration, MySQL has to be installed and the service has to be started. It can be done using the following command.

```
$ sudo yum install mysql-server

$ sudo service mysqld start
```

MySQL connector must be configured first. It connects the Hive metastore and MySQL database.

```
$ sudo yum install mysql-connector-java
$ ln -s /usr/share/java/mysql-connector-java.jar /usr/lib/hive/lib/mysql-
connector-java.jar
```

Then the database and the user are created. Finally, the metastore is configured in such a way that data gets into and out of MySQL database. The following changes should be made in the hive-site.xml file to configure the Hive metastore.

```
  <property>
  <name>javax.jdo.option.ConnectionURL</name>
  <value>jdbc:mysql://myhost/metastore</value>
  <description>the URL of the MySQL database</description>
```


Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

```
</property>

<property>
  <name>javax.jdo.option.ConnectionDriverName</name>
  <value>com.mysql.jdbc.Driver</value>
</property>

<property>
  <name>javax.jdo.option.ConnectionUserName</name>
  <value>hive</value>
</property>

<property>
  <name>javax.jdo.option.ConnectionPassword</name>
  <value>mypassword</value>
</property>

<property>
  <name>datanucleus.autoCreateSchema</name>
  <value>false</value>
</property>

<property>
  <name>datanucleus.fixedDatastore</name>
  <value>true</value>
</property>

<property>
  <name>datanucleus.autoStartMechanism</name>
  <value>SchemaTable</value>
</property>

<property>
  <name>hive.metastore.uris</name>
  <value>thrift://<n.n.n.n>:9083</value>
  <description>IP address (or fully-qualified domain name) and port of the
metastore host</description>
</property>
```

Thus the configurations are made in in the metastore so that the data can be analyzed using Hive query language.

8.2. DATA STORAGE

The streamed Twitter data are stored in HDFS. This Twitter data is never modified or altered in any situation.

As the Twitter data is in JSON format, it will not work with the default setup. Hive SerDe [9] is used to interpret the data which is one of the biggest advantage of Hive. SerDe stands for Serializer and Deserializer which helps in converting the data into format which Hive can understand.

For example, SerDe takes the following JSON format tweet.


```
{
  "filter_level":"medium",
  "contributors":null,
  "text":"@googlechrome Why won't google chrome open?",
  "geo":null,
  "retweeted":false,
  "in_reply_to_screen_name":"googlechrome",
  "truncated":false,
  "lang":"en",
  "entities":{
    "symbols":[

    ],
    "urls":[

    ],
    "hashtags":[

    ],
    "user_mentions":[
      {
```


Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

```
        "id":56505125,
        "name":"Google Chrome",
        "indices":[
          0,
          13
        ],
        "screen_name":"googlechrome",
        "id_str":"56505125"
      }
    ]
  },
  "in_reply_to_status_id_str":null,
  "id":436622096794669056,
  "source":"web",
  "in_reply_to_user_id_str":"56505125",
  "favorited":false,
  "in_reply_to_status_id":null,
  "retweet_count":0,
  "created_at":"Thu Feb 20 22:03:14 +0000 2014",
  "in_reply_to_user_id":56505125,
  "favorite_count":0,
  "id_str":"436622096794669056",
  "place":null,
  "user":{
    "location":"Corona, CA",
    "default_profile":true,
    "profile_background_tile":false,
    "statuses_count":1429,
    "lang":"en",
    "profile_link_color":"0084B4",
    "id":165185690,
    "following":null,
    "favourites_count":18,
    "protected":false,
    "profile_text_color":"333333",
    "description":"I'm a Dodger fan who loves baseball so much I have two twitter acco
unts. I'm also a music junkie.",
    "verified":false,
    "contributors_enabled":false,
    "profile_sidebar_border_color":"C0DEED",
    "name":"Jessica Navarro",
    "profile_background_color":"C0DEED",
```

```
    "created_at":"Sat Jul 10 21:24:04 +0000 2010",
    "is_translation_enabled":false,
    "default_profile_image":false,
    "followers_count":56,
    "profile_image_url_https":"https://pbs.twimg.com/profile_images/378800000616135881
/16f82ea671f7465b41f5829622cb5988_normal.jpeg",
    "geo_enabled":false,
    "profile_background_image_url":"http://abs.twimg.com/images/themes/theme1/bg.png",
    "profile_background_image_url_https":"https://abs.twimg.com/images/themes/theme1/b
g.png",
    "follow_request_sent":null,
    "url":null,
    "utc_offset":null,
    "time_zone":null,
    "notifications":null,
    "profile_use_background_image":true,
    "friends_count":137,
    "profile_sidebar_fill_color":"DDEEF6",
    "screen_name":"Dodger_Jess83",
    "id_str":"165185690",
    "profile_image_url":"http://pbs.twimg.com/profile_images/378800000616135881/16f82e
a671f7465b41f5829622cb5988_normal.jpeg",
    "listed_count":4,
    "is_translator":false
  },
  "coordinates":null
}{
  "filter_level":"medium",
  "contributors":null,
  "text":"Google Chrome just added emojis!! I've been waiting for this! ☺",
  "geo":null,
  "retweeted":false,
  "in_reply_to_screen_name":null,
  "truncated":false,
  "lang":"en",
  "entities":{
    "symbols":[

    ],
    "urls":[

    ],
```


Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

```
    "hashtags":[

    ],
    "user_mentions":[

    ]
  },
  "in_reply_to_status_id_str":null,
  "id":436622163446366208,
  "source":"web",
  "in_reply_to_user_id_str":null,
  "favorited":false,
  "in_reply_to_status_id":null,
  "retweet_count":0,
  "created_at":"Thu Feb 20 22:03:30 +0000 2014",
  "in_reply_to_user_id":null,
  "favorite_count":0,
  "id_str":"436622163446366208",
  "place":null,
  "user":{
    "location":"Los Angeles",
    "default_profile":false,
    "profile_background_tile":true,
    "statuses_count":136814,
    "lang":"en",
    "profile_link_color":"59C2C2",
    "profile_banner_url":"https://pbs.twimg.com/profile_banners/23286831/1374483591",
    "id":23286831,
    "following":null,
    "favourites_count":243,
    "protected":false,
    "profile_text_color":"2BA857",
    "description":"What's understood doesn't need to be explained.",
    "verified":false,
    "contributors_enabled":false,
    "profile_sidebar_border_color":"FFFFFF",
    "name":"Telejah Dean ♥□",
    "profile_background_color":"959C9E",
    "created_at":"Sun Mar 08 07:18:31 +0000 2009",
    "is_translation_enabled":false,
    "default_profile_image":false,
    "followers_count":1224,
```

```
    "profile_image_url_https":"https://pbs.twimg.com/profile_images/432045447717216257
/-hTZVEIQ_normal.jpeg",
    "geo_enabled":true,
    "profile_background_image_url":"http://pbs.twimg.com/profile_background_images/818
913900/12b53ccdd455fbe03e67dfcf8d239510.png",
    "profile_background_image_url_https":"https://pbs.twimg.com/profile_background_ima
ges/818913900/12b53ccdd455fbe03e67dfcf8d239510.png",
    "follow_request_sent":null,
    "url":"http://married2thamoney.tumblr.com/",
    "utc_offset":-28800,
    "time_zone":"Pacific Time (US & Canada)",
    "notifications":null,
    "profile_use_background_image":true,
    "friends_count":104,
    "profile_sidebar_fill_color":"000000",
    "screen_name":"Telejah",
    "id_str":"23286831",
    "profile_image_url":"http://pbs.twimg.com/profile_images/432045447717216257/-
hTZVEIQ_normal.jpeg",
    "listed_count":36,
    "is_translator":false
  },
  "coordinates":null
}
```


The following query converts the JSON format above into queryable format.

```
SELECT created_at, entities, text, user
FROM tweets
WHERE user.screen_name= 'Dodger_Jess83';
```

JSON format after converting into queryable format will look like,

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

created_at	entities	text	user
Thu Feb 20 22:03:14 +0000 2014	{"urls":[],"user_mentions":[{"screen_name":"googlechrome","name":"Google Chrome"}],"hashtags":[]}	@googlechrome Why won't google chrome open?	{"screen_name":"Dodger_Jess83","name":"Jessica Navarro","friends_count":137,"followers_count":56,"statuses_count":1429,"verified":false,"utc_offset":null,"time_zone":null}

8.2.1. DATABASE SCHEMA:

The Hive warehouse has three tables initially. Tweets table, Time_zone_map and dictionary table are in the first stage of the process. Time_zone_map and dictionary files are manually loaded into the hive warehouse and the tables are created. Tweets is an external table created from the data in the HDFS. A new table named tweets_country is created as the tweets table has only the time zone and not the country. It is created my joining Tweets and Time_zone_map table. The tweets table and Dictionary table are joined to produce the tweets_sentiment table which consists of tweets with their sentiment.

The diagram below is the schema of the database in Hive.

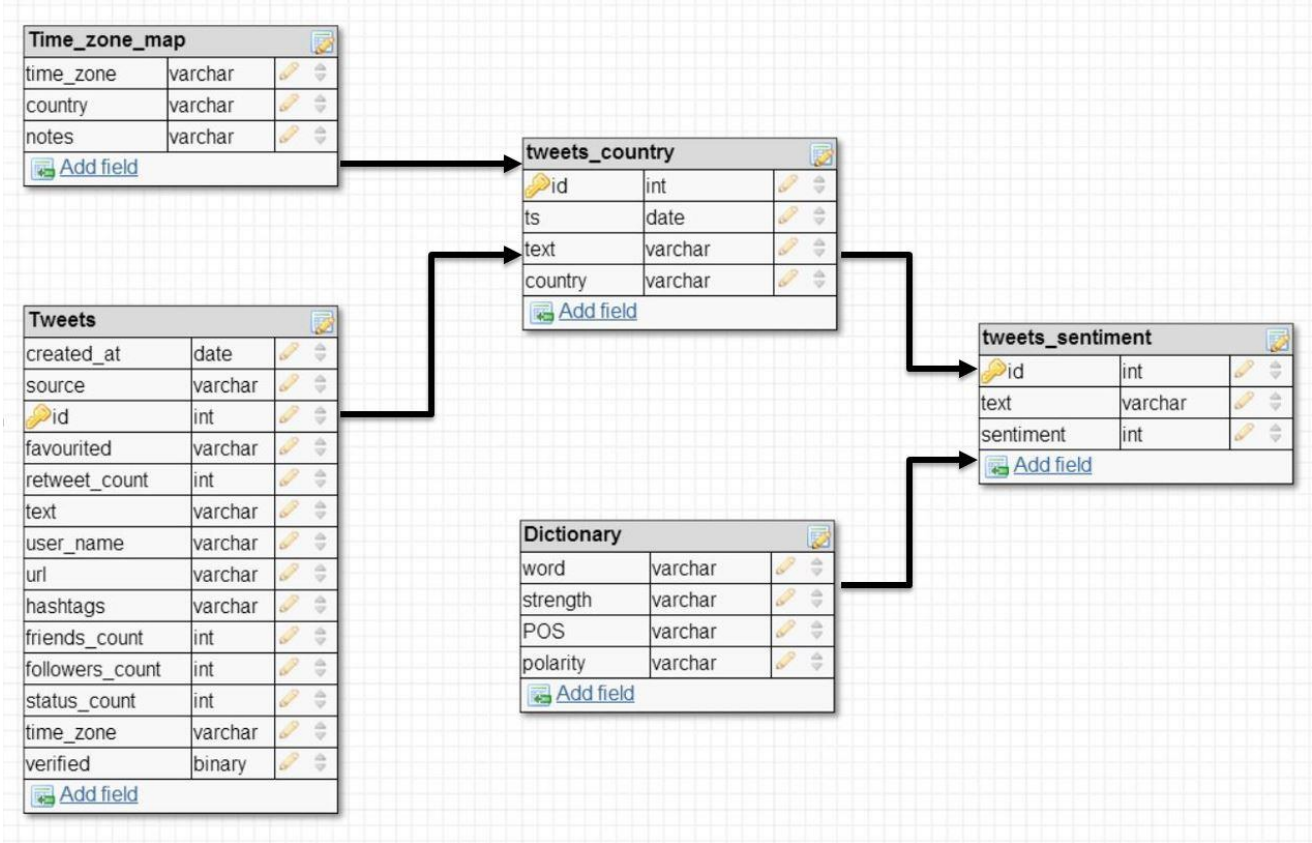


Figure 10: Database Schema

8.3. DATA FILTRATION

Classification in Filtering:

Data are classified into 2 main categories. They are,

- Polarized (Positive & Negative)
- Unpolarized (Irrelevant)

Polarized words can be further classified into positive and negative based on their usage. They are classified as positive and negative depending on the sentiments they deal with. For example,

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

words like wonderful, amazing gives a positive feel whereas words like terrible, awful gives a negative feel.

There is another set of words which do not convey any sentiment. Those are irrelevant words. In this project we deal only with positive and negative words that convey sentiment.

Data Filtering is done based on various things like emoticons, acronyms, repeated punctuations, upper case, username and hashtags.

Emoticons:

:-) , : -(, :-o, :/, …

These are some the emoticons [17]. Emoticons are the best way to express or convey an emotion in a statement. In the recent years these emoticons play a large role in every form of written communication. They play a significant role in social networking websites.

In this project we are making use of these emoticons to classify statements into positive and negative statements. For example - if Netflix crashes due to Microsoft Silverlight, the user will tend to tweet a negative statement with : -(emoticon. This means the user is frustrated and not satisfied with performance of Netflix or Microsoft Silverlight.

Similarly different emoticon deliver different emotion which we can use for our analysis process. Emoticons has one of the highest weightages in the sentiment word dictionary.

Sample Tweets	Sentiment
I am going to bed early tonight :-)	Positive
I am going to bed early tonight :-(	Negative
I am going to bed early tonight :-|	Neutral
I am going to bed early tonight :-D	Positive
I am going to bed early tonight ;-)	Positive

The table above has exactly same words in all the tweets but conveys different emotions as it has different emoticons. Thus emoticons play important role in deciding the sentiment of a sentence.

Acronyms:

LOL, OMG, ROFL, …

Recently, there is a wide usage of acronyms in social networking sites and forums. It depicts the thought of the user evidently. Even though there is some negative sense in the tweets and there is a LOL acronym in the later part of the tweet, then the entire tweet gains a positive polarity.

Acronyms have high weightage equal to the emoticons. They clearly show the sentiment of the entire tweet.

Repeated Punctuations:

!!!!! , ???? , …

Punctuation have their purpose in a statement whereas continuously repeated punctuations emphasize that something important is said. These repeated punctuations are seen in people's tweets and statuses in social networks. In this project we are taking into account of repeated punctuations as important criteria. For example – '!!!!!' conveys several meanings like extremely important, extremely urgent, terrible or amazing. So priority is given for tweets with repeated punctuations.

Upper Case Identification:

ALL_CAPS keyword (eg., REALLY)

In social networks, people tend to use ALL_CAPS keyword to emphasize the importance of something. In this project, these keywords in the tweets will be noted seriously and will be given priority similar to the repeated punctuations.

Hashtags '#' and Usernames '@':

As the entire project is based on the tweets from Twitter, hashtags and usernames must be given priority. Hashtags are used to filter the tweets on a particular topic. For example, '#Google Chrome' deals with all the tweets on the topic Google Chrome. Similarly, usernames like '@Mircosoft' might say something related to its Silverlight which might help this project. Therefore hashtags '#' and usernames '@' are given priority in the project.

8.4. DATA ANALYTICS

Data Analytics is third and important stage in this project. This stage deals with analyzing the filtered twitter data using Hive Query Language. This stage provides us the results which we want.

The data is actually stored in the external table in HDFS. The Hive accesses the data without moving them from their location. The unstructured data which is stored in the HDFS in converted into structured data in columnar pattern using JSON SerDe from the Hive interface.

8.4.1. SENTIMENT ANALYSIS DICTIONARY:

It is a very important and integral part of the data analysis procedure. The dictionary is divided into four columns namely, words, strength, parts of speech and polarity. The words in the dictionary are classified deeply and weights are provided to them based on their position in the classification.

The original dictionary has only the words and polarities in it. The improved dictionary that is used in this project has many essential additions like emoticons, acronyms, negation. Emoticons and acronyms have very high strength as they can change emotion of the entire tweets. Negations play significant role in the sentiment analysis part [19].

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

Weights - Positive words

- For weak words
  - Noun -> 2
  - Verb -> 2
  - Adjective -> 1
  - Adverb -> 1
  - Any POS -> 1
- For strong words
  - Noun -> 3
  - Verb -> 3
  - Adjective -> 2
  - Adverb -> 2
  - Any POS -> 2
- For very strong words
  - Acronym -> 4
  - Emoticons -> 4

Weights - Negative words

- For weak words
  - Noun -> -2
  - Verb -> -2
  - Adjective -> -1
  - Adverb -> -1
  - Any POS -> -1
- For strong words
  - Noun -> -3
  - Verb -> -3
  - Adjective -> -2
  - Adverb -> -2
  - Any POS -> -2
- For very strong words
  - Acronym -> -4
  - Emoticons -> -4

Neutral words are given value zero.

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

Apart from positive, negative and neutral sentiment, a new category called ‘both’ is introduced. These words can be represented as both positive and negative.

Steps involved in data analysis stage are,

- Tweets table is created and data are loaded into it.

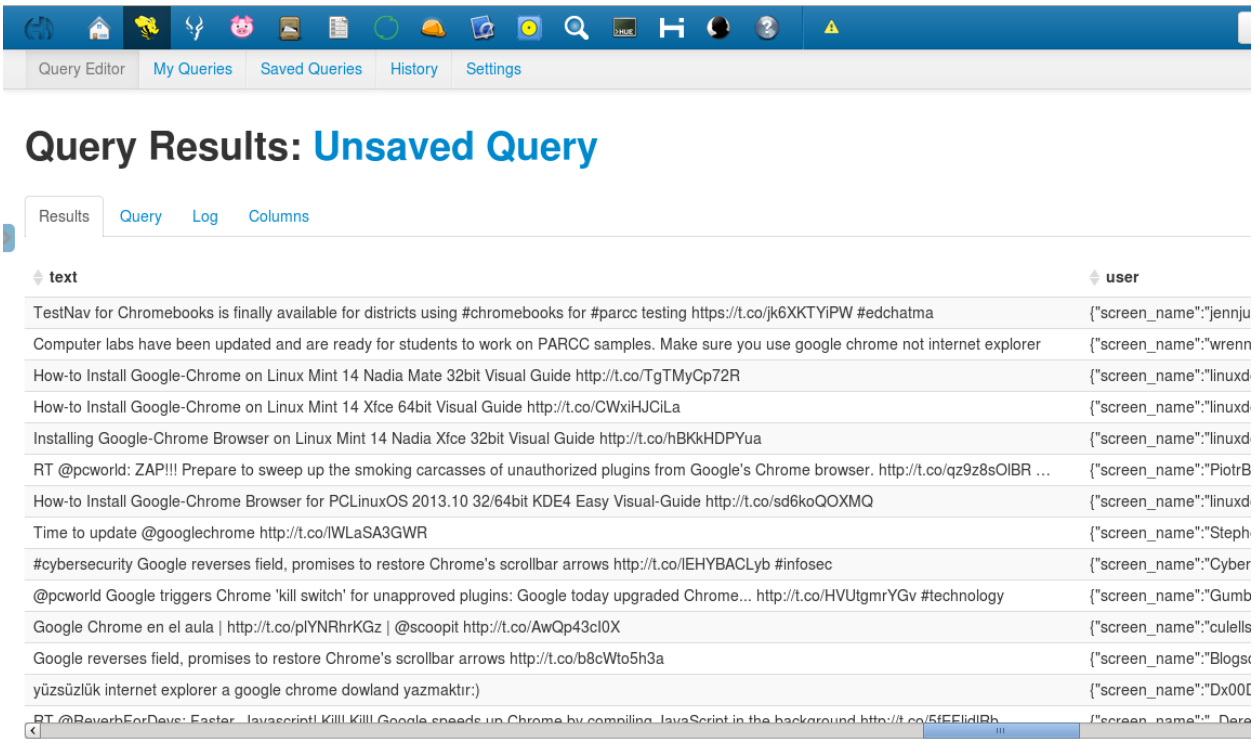


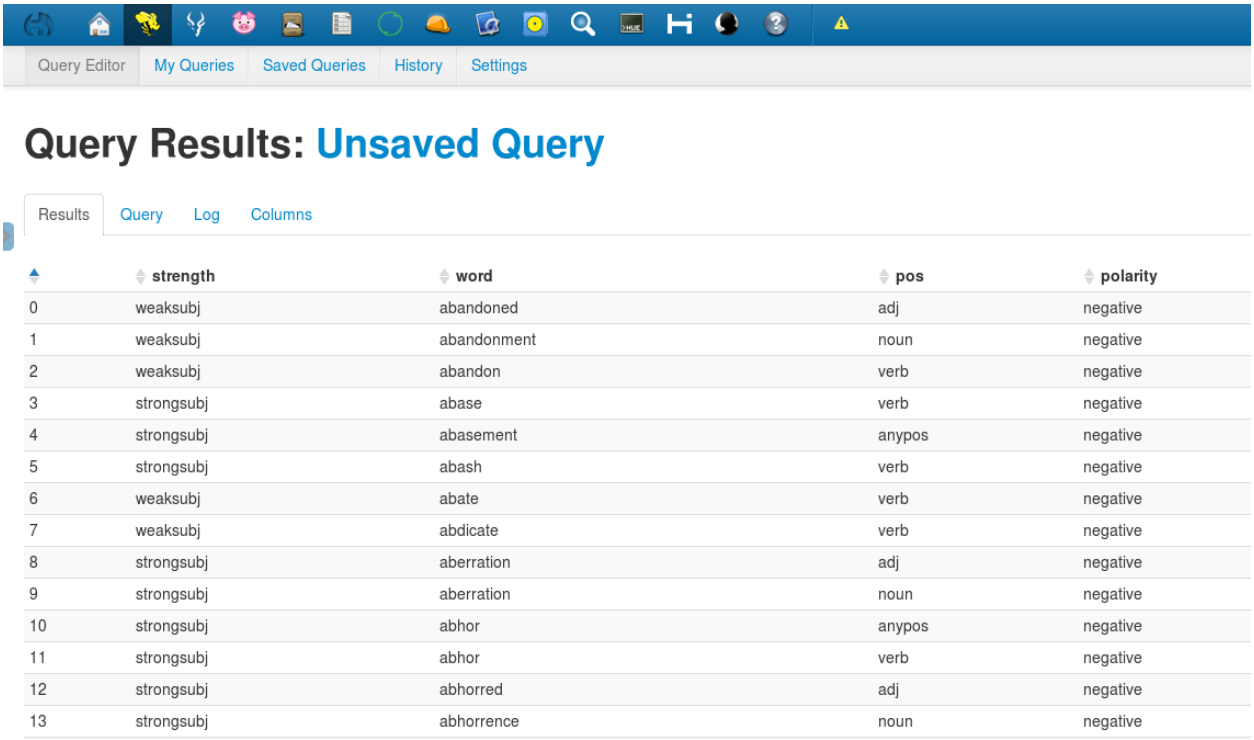
Figure 11: Segment of Tweets table

The image above in just a segment of the entire Tweets table. The structure of Tweets table will be

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

col_name	data_type
id	bigint
created_at	string
source	string
favorited	boolean
retweet_count	int
retweeted_status	struct<text:string,user:struct<screen_name:string,name:string>>
entities	struct<urls:array<struct<expanded_url:string>>,user_mentions:array<struct<screen_name:string,name:string>>,hashtags:array<struct<text:string>>>
text	string
user	struct<screen_name:string,name:string,friends_count:int,followers_count:int,statuses_count:int,verified:boolean,utc_offset:string,time_zone:string>
in_reply_to_screen_name	string

- Upload the improvised dictionary to the Hive warehouse.



	strength	word	pos	polarity
0	weaksubj	abandoned	adj	negative
1	weaksubj	abandonment	noun	negative
2	weaksubj	abandon	verb	negative
3	strongsubj	abase	verb	negative
4	strongsubj	abasement	anypos	negative
5	strongsubj	abash	verb	negative
6	weaksubj	abate	verb	negative
7	weaksubj	abdicate	verb	negative
8	strongsubj	aberration	adj	negative
9	strongsubj	aberration	noun	negative
10	strongsubj	abhor	anypos	negative
11	strongsubj	abhor	verb	negative
12	strongsubj	abhorred	adj	negative
13	strongsubj	abhorrence	noun	negative

Figure 12: Dictionary Table in Hive

The table is created for the dictionary and the data are imported into it. Four columns namely, strength, word, pos and polarity is created. The structure of the table will be

col_name	data_type
strength	string
word	string
pos	string
polarity	string

After loading data into the table, it can be used to provide polarity to words in twitter data.

- Upload the time zone map file to the Hive warehouse.

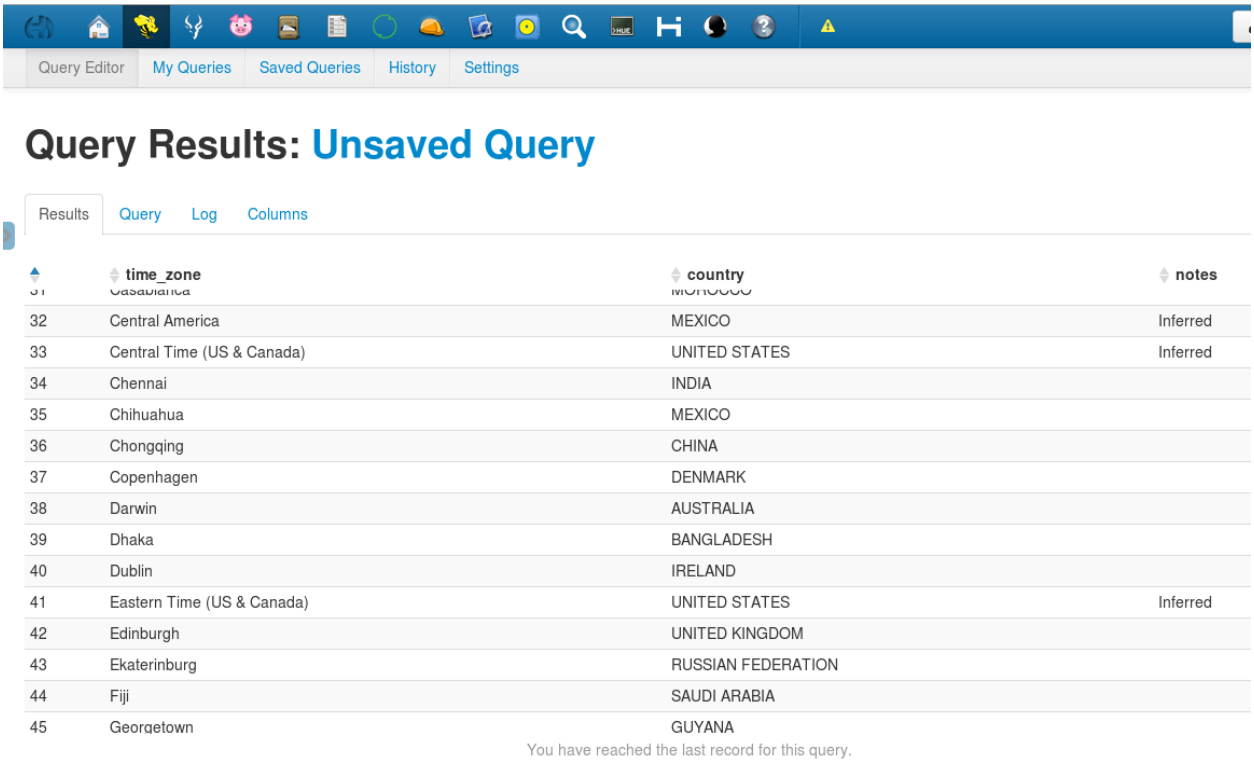


Figure 13: Time Zone Map Table in Hive

A table named time_zone_map is created and the data are loaded into the table. The structure of the table will be

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

col_name	data_type
time_zone	string
country	string
notes	string

The twitter data does not have country attribute in it. The JSON data provides just the time zone. Country attribute can be derived from the time_zone_map table by joining the tweets and time_zone_map table.

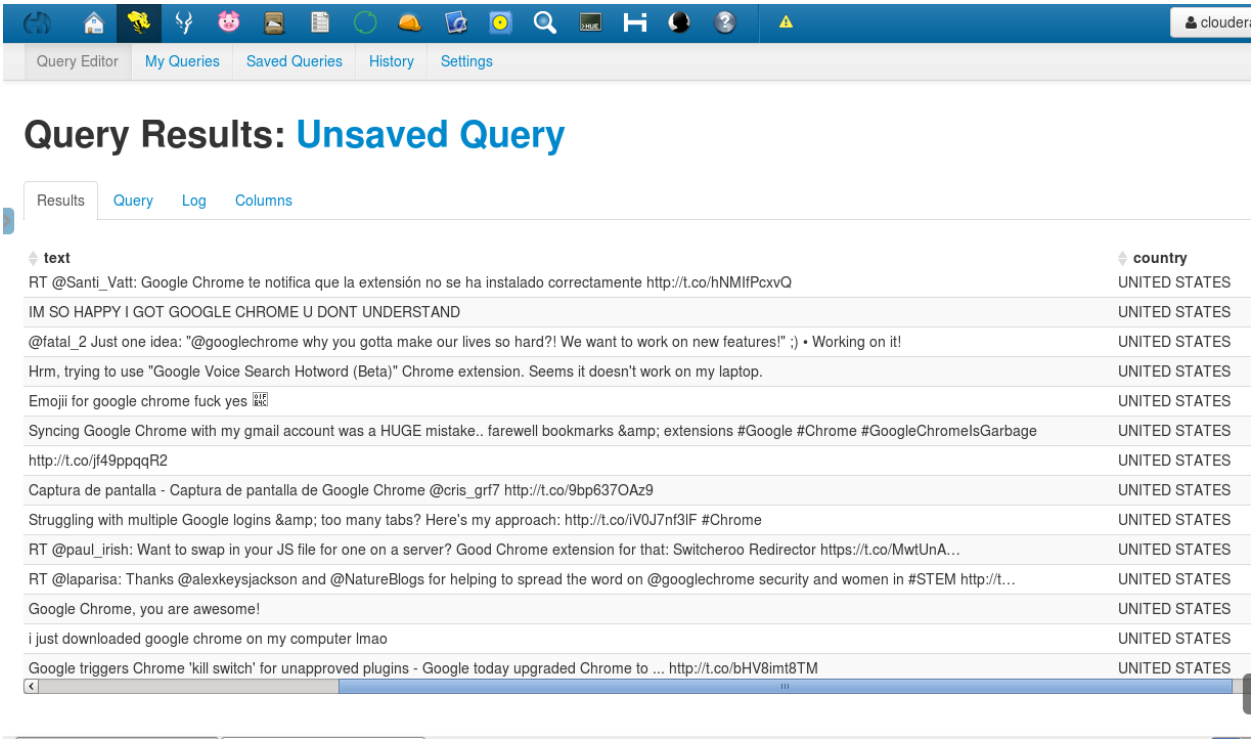


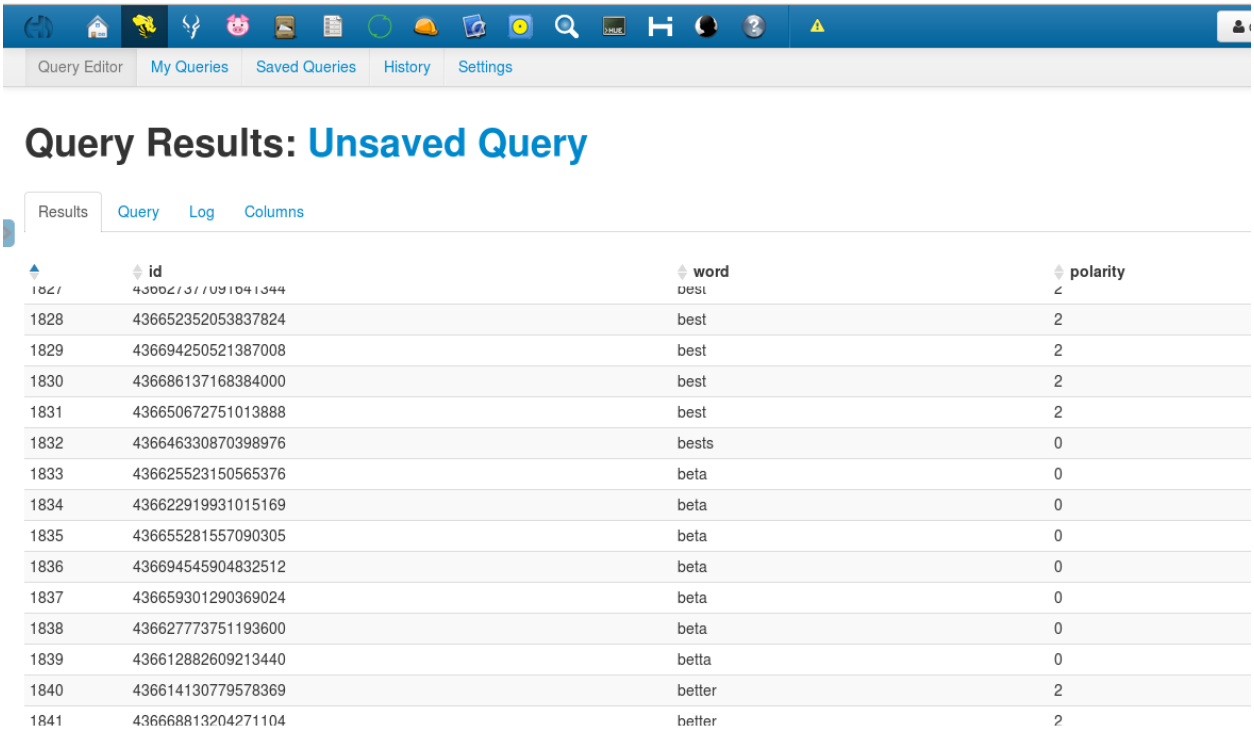
Figure 14: Tweets mapped to country

The country of the user from where the tweets id originated plays an important role in data presentation. The country location is needed in this project as problems faced by customers using Google chrome browser may be different in different countries as the bowers may

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

be updated differently based on location. There can be similar issue with Microsoft Silverlight also.

- The tweets from the Hive external table are is split up into words to make the process easy. Each and every word is compared with dictionary and provided a numerical value based on their weights and classification.



	id	word	polarity
1827	436627377091641344	best	2
1828	436652352053837824	best	2
1829	436694250521387008	best	2
1830	436686137168384000	best	2
1831	436650672751013888	best	2
1832	436646330870398976	bests	0
1833	436625523150565376	beta	0
1834	436622919931015169	beta	0
1835	436655281557090305	beta	0
1836	436694545904832512	beta	0
1837	436659301290369024	beta	0
1838	436627773751193600	beta	0
1839	436612882609213440	betta	0
1840	436614130779578369	better	2
1841	436668813204271104	better	2

Figure 15: Values assigned to words in tweets

- The words are grouped based on their ID which is the unique key in the table.
- Values of the words with same ID are summed together.
- If the result is greater than zero, then it is a positive tweets. If the result is lesser than zero, then it is a negative tweets.

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

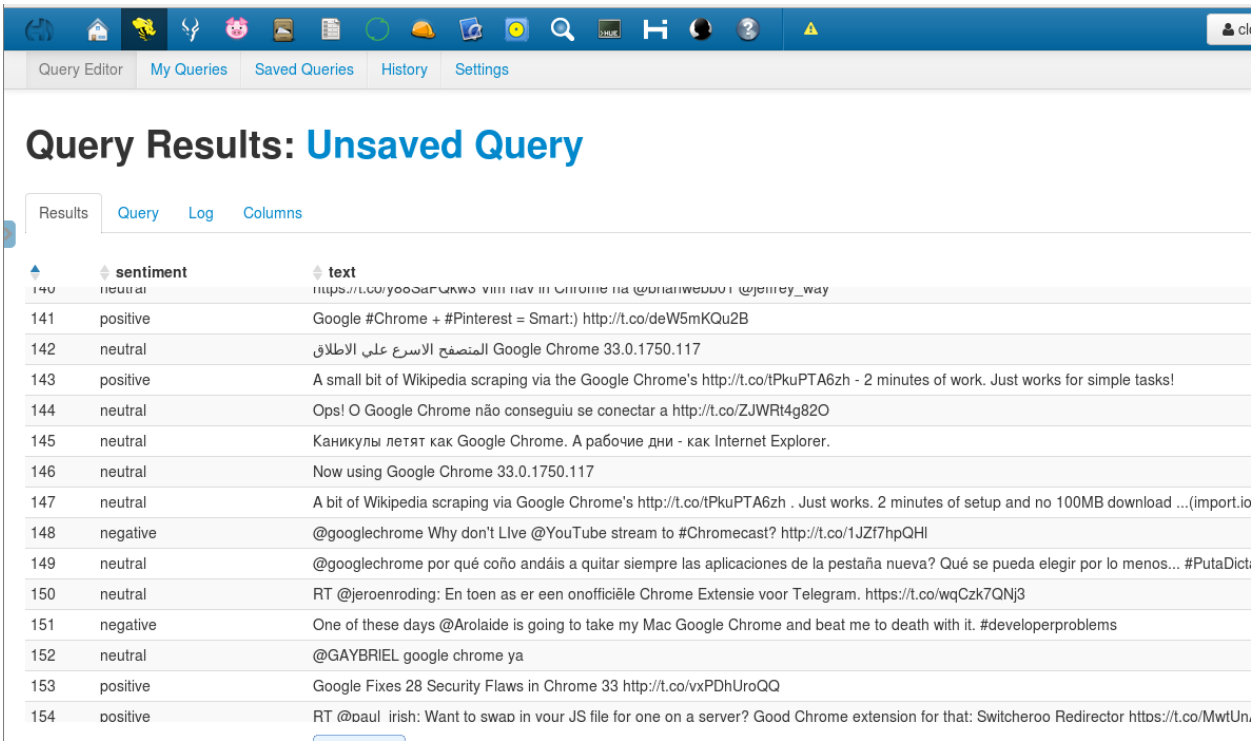
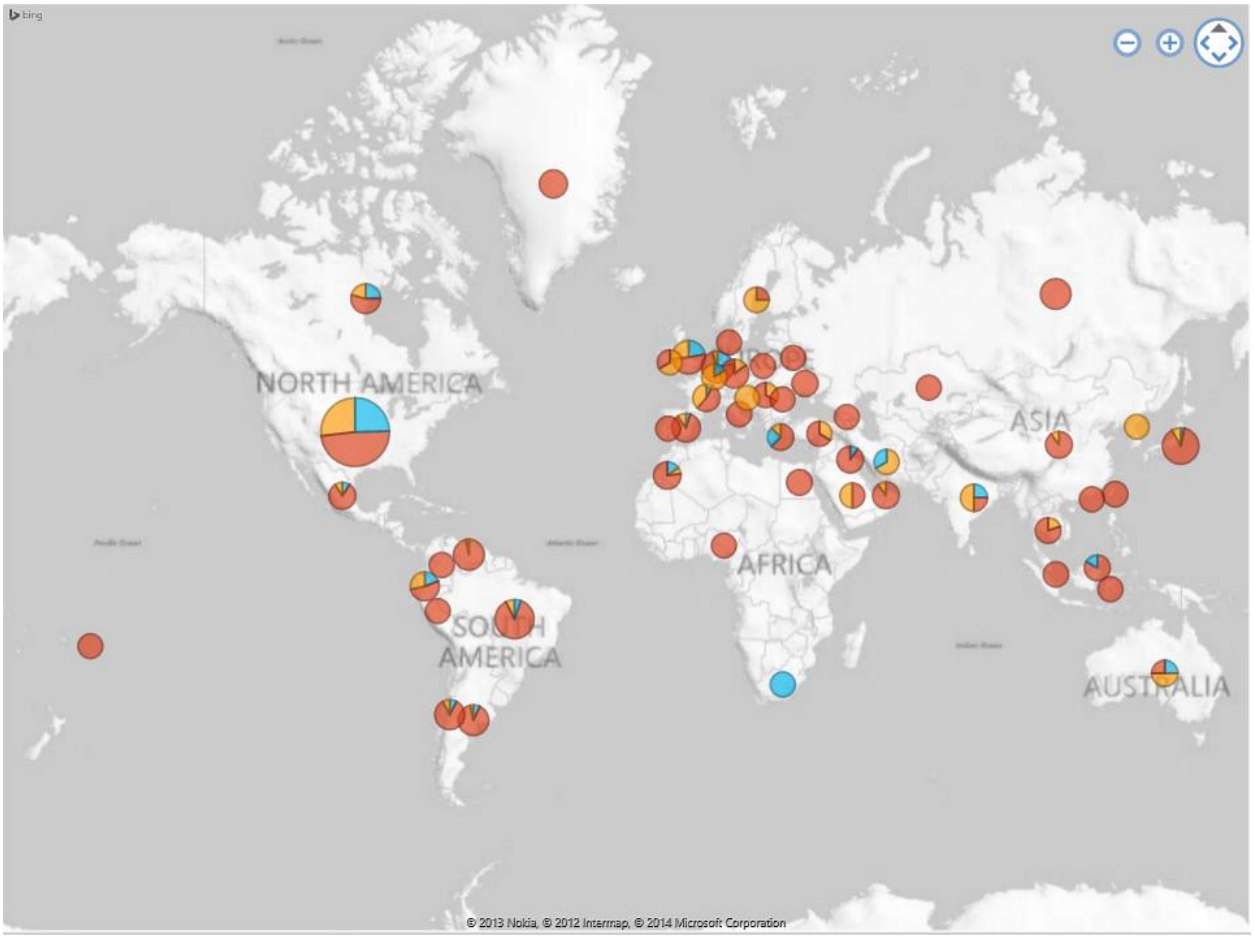


Figure 16: Sentiment of entire tweet

- Suppose the result is exactly zero, then the tweets is neutral.
- Thus the polarity is calculated for tweets from Twitter.

8.5. DATA PRESENTATION

Data Presentation is the final stage which deals with producing the results of analytics in human understandable format. Visualizations are made using the results of the analytics so that it can be easily reported to the companies if there is any problem with their free products.



Red: Positive **Blue: Negative** **Orange: Neutral**

Figure 17: Sentiments Plotted country-wise for 'Google chrome'

The image above represents the ratio or count of the positive, negative and neutral sentiments country-wise for the keyword 'Google chrome'. The Hive database in the HDFS is connected to Windows 7 operating system using ODBC connectivity. The data is imported into the Microsoft Office Excel through the ODBC connection. Country and sentiments corresponding to same tweets are selected from the final table in the database.

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

The selected data are imported into Microsoft Excel through the Power view in the insert category. Then the map view in the design category is selected to produce the desired output in the above image.

Figure 18: Google chrome's sentiment in United States

The image above shows the ratio of positive, negative and neutral sentiment for the Google chrome in the United States. Similarly the sentiment can be analyzed for any free product and visualizations can be created for the same. Form this type of visualization, a company can know the pulse of the people in various countries. This will greatly help a company to keep an eye on their products.

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

Several types of visualizations like graph or pie-chart can be created easily from the Microsoft Excel as the database is connected to Windows using ODBC. The type of visualization that can be created totally depends on the criteria on which the data has to be analyzed.

9. TESTING AND RESULTS

9.1. DATA VALIDATION

The image below shows one of the tweet streamed into the HDFS by Flume. This particular tweet is filtered so that it has the keyword 'Google chrome' in it. Thus the data streaming part of the project is validated.

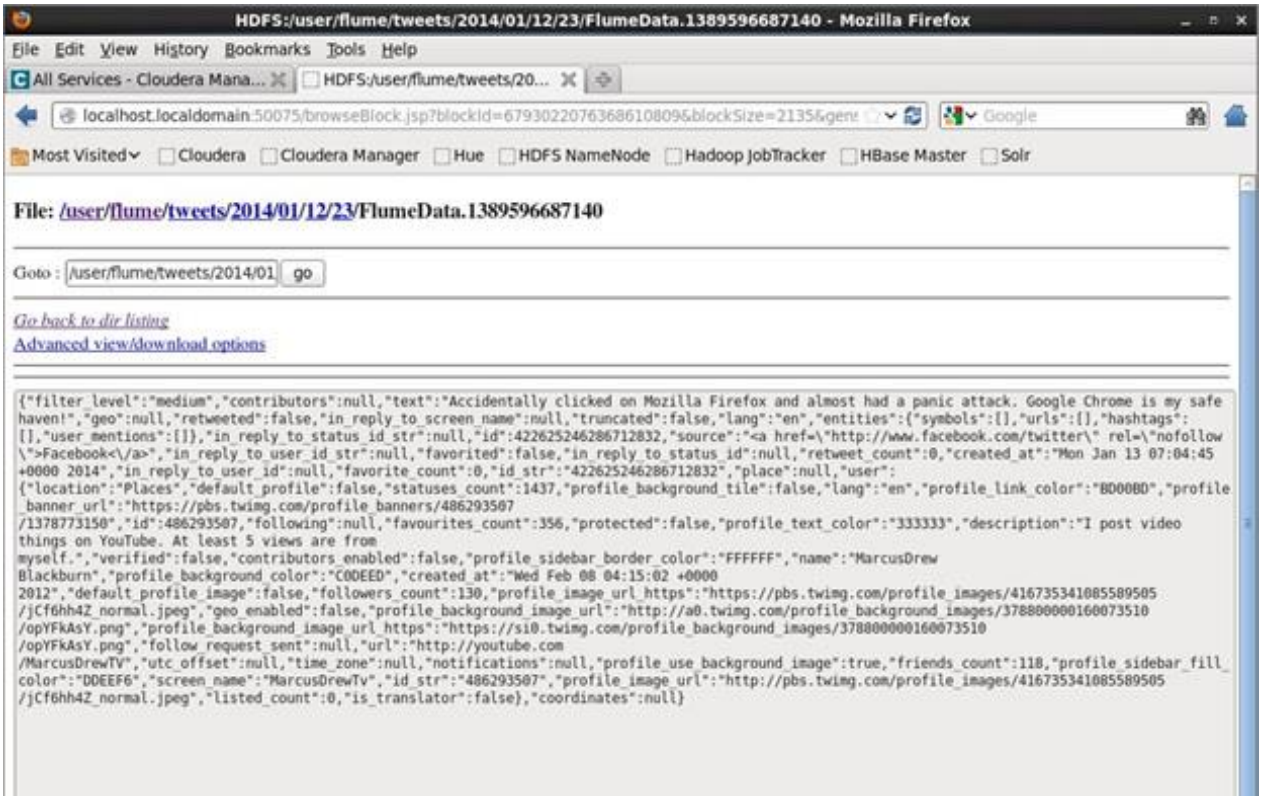


Figure 19: Twitter data in HDFS

The data analysis part is validated as per Dr.Teng Moh's instruction. A test data was created which consists of random tweets with almost every possible types of sentences. The test data contains 250 tweets on which data analysis was performed. Before performing the data analysis, the test data is manually evaluated and sentiment is provided for every tweets which will be a standard result to which the results of data analysis can be compared to. Now, the data analysis on test data is carried on by triggering the map-reduce function using HiveQL. In the map-reduce function, the words in the test data are given polarity which it turn provides sentiment to the entire tweets.

Comparison between manually calculated sentiment and sentiment computed by the tool:

my_polarity	sentiment	text
positive	positive	I am going to bed early tonight :-)
negative	negative	I am going to bed early tonight :-(
neutral	neutral	I am going to bed early tonight :-|
positive	positive	I am going to bed early tonight ;-)
positive	positive	I am going to bed early tonight :-D
negative	negative	I am going to bed early tonight :-/
neutral	neutral	I am going to bed early tonight :-P
positive	positive	I love the faces Serena makes. she always makes me laugh lol
positive	positive	The quality on some girls selfies makes me wonder if they took it with their little brothers Nintendo DS lmao
positive	positive	me and my friends are so mean to each other it makes me rofl
neutral	neutral	2moro is The big day
neutral	neutral	Voting is over. I'll go through the pics and tally up the numbers. BRB
neutral	negative	"BTW, there's a MN U.S. Senate debate on April 1. It's open to the media and all the major candidates are expected"
positive	positive	Off here meeting after lunch catch U all later mates B4N
positive	positive	Looks like the Unicorn Porthtowan beckons. Acoustic night of music in Porthtowan. BCNU
positive	positive	Happy birthday to my sweet BFF
positive	positive	YAY! I'll CYA there
negative	negative	DBEYR - Don't Believe Everything You Read
negative	negative	Through The Night DILLIGAS
negative	negative	"ya totally dude..... not reporting on day trading trends at all, just some good old FUD"

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

my_polarity	sentiment	text
neutral	neutral	Pomeroy prediction was a 62-61 Illinois win. FWIW
positive	positive	dressed like a 50 year old art teacher and I LOVE IT new tattoo new birkenstocks GR8 DAY
positive	positive	SOME OF THE AWESOME NEW MUTUALS/FRIENDS IVE MADE THIS WEEK ILY
positive	positive	IMHO she's kind of all over the place.
positive	positive	twitter people are so much better than people IRL
positive	negative	new socks weeeee! also got some w monkeys those will be previewed L8R
positive	positive	i don't even know how to use tumblr... LMAO password is changed
positive	positive	tnx u all for listening ;)
positive	positive	You really made me smile! I love all ur interviews too! Stay awesome! XOXO
positive	positive	Joke of the day.. Alagiri appreciates J Amma having maintaining decipline in her party. #EKSI
negative	negative	The hottest chick just got killed in the first episode... WTF
negative	negative	Call their customer service and rant
negative	positive	It is a poor ability.
negative	neutral	It is a bad ability
positive	positive	It is a good ability
positive	positive	It is an appreciable ability
positive	positive	It is a respectable ability
positive	positive	His madness is awesome
negative	negative	His madness is terrible
positive	positive	Save the disabled people
negative	negative	Disabled people are never taken care
negative	neutral	He bragged that he had won
negative	negative	Laughin so hard because Michelle hates rats and I'm sending her disgusting google images #dying
positive	negative	"Anytime I'm upset, I google ""Britney Spears ugly face"" and my mood immediately improves."
negative	negative	Your latest version keeps crashing in every browser I've got. Uninstalled and tried again and it's still crap.
negative	negative	Hello @Silverlight - could you update your plugin so it doesn't crash constantly in all browsers on Mac? Makes @netflix unusable. Thanks!
negative	negative	I don't know what Microsoft silverlight is but it's pissing me off.
neutral	positive	Blind love sentiment ends long after intellectual reasons are exhausted. Why does it have to take so so long?
negative	negative	I can't think of anyone except Julie Larson Green as the single person who has most negatively impacted Microsoft's culture and its products
negative	negative	"Fuck Xbox tho, i havent touched that shit in a long time...its called growing up i guess "
neutral	positive	YouTube has led me down many paths... none of which have anything to do with the hw I'm doing lol.

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

my_polarity	sentiment	text
negative	negative	But my data is more than 5TB! Your life now sucks - you are stuck with Hadoop.
neutral	negative	"In the strange world of Big Data, Luigi and Kafka ride together on a yellow toy elephant who's name is Hadoop. "
negative	negative	Just had a cough attack and liked this really annoying guys Facebook status on accident ugh
negative	negative	I swear police are not helpful what's so ever
positive	positive	I saw a guy wearing a Quora hoodie in San Jose today. I love being back in the Bay (*ï¿½?ï¿½*)
negative	negative	Really regret buying this @lenovo. Haven't had it 3 weeks and it constantly freezes and now will go to a black screen without sleeping.
positive	negative	I farted in the Apple store the other day and everyone got pissed. Not my fault they don't have Windows.? This got me Lol-ing too
positive	negative	"I seriously wanna cry, I'm so happy my favorite person in the world is finally outta the hospital! ?????????? "
positive	positive	That?s because Stanford is awesome and definitely cooler than UC Berkeley.
negative	negative	He was at SJSU practice last week. He nearly broke my hand when he shook it.
positive	negative	"Yeah, sjsu is on spring break so it's not as crowded, enjoy it -stranger at Starbucks ????"
negative	negative	Only my mom would plan a 8am doctors appointment on spring break...
negative	negative	It is with deep sadness and regret that I learned the announcement on the lost of all passengers of #MH370. *SBY*
negative	negative	"My deep condolences are to their families, and my prayers to those on #MH370. May the soul of their loved ones rest in pea?"
positive	positive	I will admit that T20 as a format is really growing on me. I think I actually like it *washes mouth with soap*
positive	positive	World T20: Team India has plenty to smile about - The Times of India http://t.co/bMcj51kvBU
positive	positive	Life is what you decide to celebrate from sunrise to sunset?.have a happy Holi everyone.
positive	positive	"@psychologicaI: The less you reply to negative people, the more peaceful your life becomes."
positive	negative	I hate when people get on Facebook or Instagram asking who wants to FaceTime ??. Bitch find some friends lol
positive	positive	Facebook and Oculus merge could be incredible. Imagine if every status update created a little virtual world that could ?
positive	positive	"He was freaking amazing in the titanic, great in Gatsby, I can't think of a bad Leo movie."
negative	negative	"The question is, do I want to go to Walgreens to get one of those crappy machine coffees?!?.."
negative	negative	"After the lost plane, what will CNN focus on next for ratings? CNN has turned into the Walmart of News "

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

my_polarity	sentiment	text
negative	negative	Walmart has no shame. Walmart relies upon govt programs in order to keep its wages low and now this? Walmart sucks.
neutral	neutral	Hortonworks Sandbox - portable Hadoop environment that comes with a dozen interactive Hadoop tutorials http://t.co/ug4bX67C6T
positive	positive	"I love that feeling of being in love, the effect of having butterflies when you wake up in the morning. That is special. Jennifer Aniston "
positive	negative	Talking to daddy about straight a's and top schools while songs blasting away. Feeling happy cause daddy not angry at me anymore.
positive	positive	"Saw an add on the subway that said ""happy xiumin day"" so cute"
negative	negative	I was supposed to go to Santa Cruz this weekend but the weather screwed it up??
negative	negative	I hate that I fought with my mom. It ruined my plans to go to Half Moon Bay to see my twin
neutral	negative	I literally can't wait to go to the boardwalk and get Tyler another hermit crab named skvid ????
negative	negative	I hate when you can't sleep and you just lay in bed and think.
negative	negative	"It's 2am, and now I can't sleep because the #Knicks have made me so angry"
negative	negative	what if I die in my sleep!? @hoopermindset would be so upset that he ignored me
neutral	negative	This has been some of the most unusual sleeping I've had today and I am going back to sleep in like an hour or two
positive	positive	"Congratulations to our new executive board Cullen Tyndall, Andrea Nolasco, Drake Phillips, and Anjali Venkat. It's going to be a great year! "
positive	positive	krish better bring my calculator to english tomorrow
positive	positive	"You're like a Shahid Afridi's sixer. Whenever you hit it, you make me go like boom boom! "
neutral	neutral	"If Dhawan is New Sehwag, Pujara is New Dravid, Virat is New Sachin Then I am New Einstein bcoz I had got 25 marks in Science in 6th class..!"
positive	positive	"Nash needs to get google glasses and YouTube his life, I'd die watching those videos lmao "
negative	negative	I'VE NEVER WANTED TO SEE SHAILENE WOODLEY PUNCH KATE WINSLET IN THE FACE SO BAD
positive	positive	Pleasantly entertained all the way through. Slightly edgier in tone compared to HUNGER GAMES. Woodley & James got dat chemistry.
positive	positive	"Hi. Beautiful clear sky over Chopin Airport. 5 degrees, wind N 10km/h. No rain today, temp.max 13 deges. "
neutral	neutral	When I move out I'm going to find a place next to both a chipotle and a wingstop.... And a baskin robins or Starbucks.
positive	positive	Oh my god I'm American and I love Starbucks and Chipotle!
negative	positive	I know it's not good to eat maggi for almost everyday but that's the only food I have.
positive	positive	I guess that's the beauty of a serendipitous moment. The endless exuberant aura spreading through each part of us when we ~ @FableEllis
positive	positive	"finally gonna go see Divergent tomorrow, hope it does the book justice "

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

my_polarity	sentiment	text
negative	negative	"You call this online magazine home page? Vikatan online UX is crap. No, I am not subscribing again"
negative	positive	"All those who send me #Farmville request on @Facebook, I can't honor such request. I actually have a real job. Wonder what you guys do. "
negative	negative	Can't understand why Facebook would buy oculus rift... are they going to make a farmville version for it?
positive	positive	"Downey is the group leader, keeping the cast intact after tough renegotiations with Marvel last year..."
neutral	neutral	i keep thinking i watched the avengers like last week but it was probably like a month ago
negative	positive	Well this is the worst i've felt in a long time. Done a quick google search and i've definitely got SARs... Just my luck.
neutral	neutral	Twitter now lets you add 4 photos to one tweet and tag anyone in them
negative	negative	"Five firms failed the Fed stress tests, but BofA and Goldman came close as well. They had to retake the test:"
neutral	neutral	"What happens when your parachute fails at 97,000 ft? Catch ""Chase in Space"" presentation at 3 o'clock tomorrow! "
positive	positive	Everytime I use my light switch I admire it's 45 degree chamfer ?? I don't even know who I am anymore ??
negative	negative	Hashtag LT Hate it but its better than living with my parents and working for a 300 dollor paycheck or not at all
positive	positive	"I'll buy it!! 10,000 yen!!!! Good price!!! "
positive	positive	My chunk and my boy broke my bank account today. It's all good though cause I love my babies.
negative	negative	Hey @LEGO_Group - why can't us girls make our own mini-figures in store? Only boy parts available. Disappointing.
neutral	neutral	I kinda feel bad about selling these Lego tickets but I think an indoor water park will be funnier
neutral	negative	"People keep telling me I look tan, but for break I went to Fort Wayne and watched Netflix 80% of the time soooo...."
negative	negative	I firmly believe lil wayne is a horrible artist
positive	positive	Netflix + couch + warm blankets + Edy's Red Velvet Ice Cream (in Lucy's princess bowl) = perfect
neutral	neutral	Just letting you all know that American Psycho is on Netflix
negative	positive	But Hershman hasn't fixed anything at HBO in the areas he was expected to. Fights like Golvokin-Chavez being on PPV don't help either
negative	positive	I'm usually a pretty outgoing person but if I'm around a girl I really like I typically forget how to speak English.
negative	neutral	Power is officially out in here in west #bridgetown and probably wont b back until morning #juanabe #noreaster #novascotia #maritimes
negative	negative	just thought my electric blanket was broken. I swear I just had a heart attack
negative	negative	Omg this movie sucks I can't do it. My heart won't allow me to betray the real Spiderman
positive	positive	"Thor 3 is confirmed, I'M SO HAPPY. "

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

my_polarity	sentiment	text
negative	negative	Wtf is this rambo? I'm all looking out for john j and I'm hearing english guys and its a 2008 ting .. I'm gonna. Watch but it looks shit
neutral	neutral	"lso Deron Williams will play but he didn't practice today. Brooklyn has had the Bobcats number of late, but this is a big game for both "
negative	negative	daily show correspondent jessica williams tweeted me. she called me bb. BB. I?M DEAD. BYE
negative	negative	"O Allah, do not give me in excess lest I may be disobedient to You, and do not give me less lest I may forget You."" Umar Ibn Khattab "
negative	negative	@cnn i lost my keys can you make a 24/7 news cycle out of that
neutral	neutral	"Need to run garbage collection so your all flash storage system won't perform like arse? Oh wait, XtremIO doesn't need to do that. "
negative	negative	"I'll never go broke. fuck bein a flash in the pan, I am a passionate man, planning with cash in my hand "
positive	positive	i think i watched the reggie jackson movie in seattle or s/t im p sure i was on vacation and saw it in theaters w my whole family
neutral	neutral	I just want to hang out with you and read comics and talk about X-Men. That's all. Come visit Seattle.
positive	positive	I haven't seen your ass in years lol come up to Seattle one day after you're back from Hawaii!
positive	positive	@JennyLeeSilver awesome! We saw them in Seattle. We're huge fans and our son LOVES him!
neutral	positive	Stylist new to Seattle? Clients needed? TCC members are looking for the ideal stylist match. We will make that match for you!
positive	positive	Met with Puducherry Chief Minister Shri N Rangaswamy in Chennai today who decided to join the NDA. I welcome him into ?
negative	negative	Srini is definitely not sitting idle in Chennai and certainly won't take things lying down. There is bound to be pressure on the judges
negative	neutral	Sleepless in Chennai. Counting Indian sheep to no avail. Still wide eyed
neutral	neutral	my dad is leaving to dubai tomorrow and joey essex is in his hotel is that good or bad
negative	negative	So jealous of my dad right now. He's in Morocco then he's going to Dubai tmr.
positive	positive	I honestly hope the believe movie hits number one on itunes it deserves so much more recognition than it gets it's truly inspiring
positive	positive	I hope the believe movie hits #1 on the iTunes chart it really deserves so much recognition it's such a great and insp?
positive	positive	You know you're living in the future when you set your iPhone Touch ID to work with your toe. Very useful #technology #technowin
neutral	negative	So the exact reason that I don't use my iPhone as my calendar just happened- no record of my life prior to one month ago. How do I fix this?
positive	positive	"If you value me as your customer as u keep telling me you do, then you will just give me my $170 that I deserve to get."
positive	positive	Twitter is a free social networking microblogging service that allows twitter members to broadcast short posts called tweets

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

my_polarity	sentiment	text
negative	negative	I'm about to just drive out to California because I cannot take New York anymore.
positive	positive	I feel like that not many seniors went to New York. Hell I would love to be able to go and take their place for them.
positive	positive	"Again, the reason why Nate Silver is famous is that he correctly interpreted aggregate polling data when others were ignoring it."
neutral	neutral	"When you're done cleansing and analyzing your data, try to get someone with domain knowledge to look it over "
neutral	positive	"Homework in every subject, test tomorrow, quiz tomorrow.. Can we just go back to elementary school where we learned how to say the ABC's. "
positive	positive	SO EXCITED to get a beautiful frame for this print that just arrived from @stephcreekmur
neutral	positive	All of the Nigerian devs who think they are too good for WP seem to like bootstrap. Prolly why their sites end up looking al?
negative	negative	One of the cats stole the piece of cooked veal I had cooling on the counter. Half eaten. On the floor. Couldn't even finish it. Bastards.
negative	negative	"Why do people post duck face selfies on insta saying ""had a terrible day"". Are they like, oh my day's so bad I need to take a? "
positive	positive	Girls should smile more often! More attractive then the duck face
negative	negative	You know you're a loser when you work and spend your free time at java makers... Go home dude.
positive	positive	"an evening of talking JavaScript to talented devs, I feel assured I?m on track with my understanding of JS but there is so much more to know"
negative	positive	"bash scripting is very hard. i enjoy doing it, because it's a lot of fun, but it's a lot to get your head around. not as easy as python."
negative	negative	"Spent the first half of the morning in RPM dependency hell. Second half in Perl CPAN dependency hell. Eff this, I'm getting lunch. "
neutral	positive	"The Twitter API now requires SSL. This is a Good Thing, but broke my auto-tweet scripts. Upgrading Net::Twitter fixed."
positive	positive	"True friends are the ones who have seen all your crazy sides and still love you, just the way you are. "
positive	positive	You can preorder the Galaxy S5 on all carriers starting tomorrow at RadioShack! But wait there's more! Save $50 when you preorder!!!
negative	negative	Like 90% of charity's are probably fraud....you're just trying to do some good and they're just taking your money??
negative	negative	The shit I seen in between yesterday and today? Having hoes aint what it is no more. U better find one and get this money.
negative	negative	"Dear apple, Samsung, et al. There is no reason our keyboards need to make noise, except to irritate the world."
positive	positive	If all goes according to my grand plan I'll be joining @noahsussman in Bangalore India for the Selenium conference 2014 in Sept. Yay!
negative	negative	"Ugh why is this so difficult. I don't want a fucking credit card, why can't places accept visa debit. Shit"
positive	positive	"She's the moon to my shine, the whiskey to my water"

my_polarity	sentiment	text
neutral	neutral	My dad was a research physicist. He worked in solid state physics. Most of his work was for NASA and the moon landing.
neutral	negative	"@cloudera: #HBaseCon 2014 is coming to SF on May 5! For #Apache #HBase users and fans, missing it is unthinkable"
positive	positive	Why did I wait so late to start my ap euro project lol
positive	positive	"I can already see that my lil brother has great taste in girls, thank God! "
positive	positive	Photoset: Write away! Poetic Oracle has your words for today?s poetry and story writing. Have fun!
neutral	positive	"my voice is girly when I talk to strangers but when I?m with friends I turn into morgan freeman"" "
neutral	neutral	if you went back in time you would make the same decisions over again because that's what made you who you are today
neutral	neutral	well i just witnessed someone woman crush wednesday themselves so im just about done with today
positive	positive	cool i can try it out later today thank you
neutral	neutral	"One of the guys brought muffins to work today and the number of times I say ""that's what she said"" just reached biblical pr? "
positive	negative	"Blogging about using MongoDB: Sounds cool now, will look like hubris upon scaling failure "
positive	positive	"you're welcome, cool service. a couple of websites and Azure mobile web services. Mostly experimental for now. Using mongoDB also. "
positive	positive	"just literally love the animation on this movie, their moves are perfect, their expressions, Toothless's flying! #DayOfDragons AMAZING! "
positive	positive	As lovers of cinematography it's hard to get behind animation (no cameras). This explains why we loved WALL-E so much
negative	positive	the day you walk into a party and see people openly doing cocaine is such a moment of lost innocence. That D.A.R.E shirt waiting back home.
positive	positive	I love my Navy shirt . Can't wait to actually join & rock the uniform (:
positive	positive	i bought a sherlock holmes book and a cat book at barnes and noble today surprise
negative	negative	Every other slide during this RA class has been a graph but they aren't showing supply and demand so I'm very confused! #EconMajorProbs
negative	negative	"Sun of a gun, I forgot my graph paper in school??"
neutral	neutral	God knows the only thing keeping me from punching this man in his throat this morning was the fact that I have to go to college.
positive	positive	So many athletes doing well in college we are getting back to being Pike U! Producing great student athletes!!!
positive	positive	Oculus looks so good I wish I had friends who would watch scary movies with me
neutral	negative	Not a fan of salt water or having to dump my face in it??
positive	positive	pasta and butter is good. better with some fresh grated parm.
positive	positive	My friends parents watch my prank videos and it makes me so happy
negative	negative	One day is warm out and out of no were it gets cold. Im tired of this shit man .
negative	negative	Setting my alarm on a night genuinely makes me feel sick

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

my_polarity	sentiment	text
positive	positive	"You can turn your dreams into reality. It's just much easier to hit snooze & let the dream continue, rather than wake up and work towards it "
positive	positive	Can you believe these 3D drawings? They're so good. #9 is hilarious! see them
negative	negative	My brothers making me watch this weird cartoon called Steven universe. Cartoon network sucks
positive	positive	"If you're a junior guy, this is your only opportunity to become a sweetheart for triangle two! Take advantage of it"
negative	negative	"I once pissed my pants in first grade, wasn't my fault we were square dancing and the dyke gym teacher wouldn't let me leave "
negative	negative	soooo bad at arts/crafts my replica of Tiananmen Square looks horrid
negative	neutral	"I take a drag at the square, I feel anxious, spit dangerous "
negative	negative	"just know he got my blood fucked up, called him a ""square"" but was the main ?? when shit got real! Oh fuck no "
neutral	negative	I can throw my own TRIP at Mile Square Park with some flashlights and hoes
neutral	neutral	"anyone wanna buy a iPod nano pink the touch screen small square style to old bigger square, or the skinny rectangle one, or a broken touch??"
positive	positive	Feel free to express yourself & lets have some fun w/ your friend to @Jackpot9_Bdg Paskal Hyper Square B39 Bdg
negative	negative	Man y'all don't know how mad I would be if somebody sold me out for my dance moves
positive	positive	"I maybe Hispanic, but I learned how to square and line dance when I was little. Thanks to my music and gym teachers."
negative	neutral	I wish my mom woulda pushed me harder in dance instead of just sitting back and let me fuck around
positive	positive	"LOVE The only story that matters: Flirt with Juliet, dance with a crush. Find someone irresistible to make you crazy in love "
negative	negative	too busy for these bruised toes I've got dance class to attend to
positive	neutral	I want to take dance classes..
positive	positive	So incredibly thankful to have advanced on to the next round of OAP with Smith today. I couldn't ask for a better dire?
positive	positive	Always nice to have our friend Nigel round to play
negative	negative	Finished with another round for this week's bills. It's never ending. And pretty ridiculous. My insurance blows.
negative	negative	"well, why don't we pick up a dictionary and I'll prove you wrong!"
negative	negative	I'm so stupid that I even accidentally deleted a word off my keyboard dictionary
negative	negative	"Just saying, if you're going to type like you don't know how to spell, I'm going to throw a damn dictionary at you. "
positive	positive	What's great about the Grinnellians is that they will cheer you on when you tell them you are doing brave and terrifying things.
positive	positive	My friends are telling me that he will notice you if you'll be brave enough to talk to him
positive	positive	galileo was really brave

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

my_polarity	sentiment	text
positive	positive	"I don't want to be just one thing. I want to be brave, I want to be selfless, and intelligent, and honest, and kind. -Tobias ?? "
positive	positive	I'm having too much fun with this cause girls who act stupid think I'm glad
positive	positive	Looking for the perfect gift? We can help you and you can even ship it anywhere it needs to go!
negative	negative	Fremont softball apparel is available! Let me know if you want a sick Tshirt
positive	positive	a mini shopping spree at American Apparel would be amazing
positive	positive	Finally finished the code for my scheduler. Hopefully my algorithm is actually the most optimal.
positive	neutral	I like it. Maybe I'll encode my music with an algorithm that can only be translated with a specific key...Sell each track for $10k.
neutral	neutral	Google receives US patent for Panda algorithm
positive	positive	"congrats Syeeda! You're going to love it here, best school in the state if not the mid Atlantic region! "
negative	negative	"Titanic sinks in Atlantic, May 1912"
negative	negative	Telling her I don't feel like checking rock I'm not going do it
neutral	neutral	I find my self listing to a lot more rock again.
positive	negative	I like being busy yea it takes a lot out of me but it's cool that I'm not just sitting at home doing nothing
negative	positive	Don't disrespect me and try to be cool with me afterwards and then extend your hand for help. I'll just spit in it.
negative	negative	Intel to ditch its Hadoop software and support Cloudera instead: Intel is set to cause
positive	positive	We will always reply as soon as we can but feel free to pop your query into a DM and we can help here
negative	negative	"Not currently relevant, but this shit is so accurate it's scary"
positive	positive	"Gonna have a glass of Bombay on the rocks and watch ""The Conjuring"" in the pitch black of course lol "
neutral	positive	Life is like riding a bicycle. To keep your balance you must keep moving.
positive	positive	hi gals. What a story. We would be happy to help your food vendors out at our market on the 12th hello@hawkersmarket.com
negative	negative	I'm literally so sad I couldn't go to the zumiez interview...I'm gonna be stuck at Safeway forever
negative	negative	It's so dumb that I can't walk around with a 21 year old buying alc in Safeway
negative	negative	"i went to the doctor to talk about my kidneys, walked out with a cold. no more seeing doctors in hospitals "
negative	negative	i still have the most random shit i happened to buy when we were driving from vermont to california
negative	neutral	"Due to Richard Sherman's remarks, a communications major from Stanford now means absolutely nothing to this world. "
positive	positive	It would be a dream come true if I could go to college at the University of Florida..??
positive	positive	We'll live happily ever trapped if you just save my life. Run and tell all the angels that everything is alright.

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

The results produced by the data analysis is compared with the standard result produced manually. Totally, there are thirty one mismatch among the results which implies that 211 out of 250 were showing correct sentiment. Thus the data analysis part of the project is validated.

10.CONCLUSION AND FUTURE WORK

In this report we have discussed in details about all the stages in this project. The goal of this project to analyze the unstructured twitter data in JSON format to provide results of free products to corresponding companies. Initially, the Hadoop ecosystem is configured to perform all the functions involved in this project.

The first stage deals with streaming real-time Twitter data into the HDFS. This is performed by Apache Flume. The Twitter data which is streamed is unstructured and it is handles by Hive and analyzed using Hive-QL.

Second stage is data storage and the third stage is filtering the data which deals about various filtering techniques which will provide much concentrated data so that the result for the analytics will be more close to accurate. Various techniques are discussed in details in a separate section in this report.

Fourth stage is the data analysis part which provides positive, negative or neutral sentiment to each and every tweets which are streamed into the HDFS in the previous stage.

The future scope of this project will be to improve the sentimental word dictionary in way that the dictionary automatically adds new words from tweets that are not present in the dictionary. The dictionary will learn continuously and refine itself.

Initially, it looks for words that are not present in dictionary. There will be various types of words including noise like conjunctions, articles and URLs. They must be removed from the list. Words like conjunctions, articles and prepositions can be removed using a file containing stop words. Noise and unwanted data are common in data streamed from online sources. Removing URLs and random things like 'hs6t9op3' will be challenging as they will not be in any stop words list. The words that are not in dictionary and stop-words list will be listed in a text file and they are given the sentiment of the entire tweet.

For example, let us assume that the word 'promotion' from the tweet 'I got promotion yesterday :-(' is not in the dictionary and stop-words list. Now, we can add the word 'promotion' to a new file with a negative polarity. Contrarily, promotion is a positive word and it has been provided with wrong polarity. This problem can be solved when the volume of data increases. When we analyze billions and trillions of tweets there will definitely we several repetitions of the word 'promotion' in them. The probability of 'promotion' in a positive tweet will be obviously high. Therefore we overwrite the words in the file by adding or subtracting the corresponding numerical values to polarity. The threshold will be decided after going through lot of test data.

As a result of this process we can create a much improved sentimental word dictionary which will continuously learn and refine itself.

Lokmanyathilak Govindan Sankar Selvan
ID: 008335483

11. REFERENCES

[1] Opinion Mining. Retrieved on 9th April 2014 from the website: http://en.wikipedia.org/wiki/Sentiment_analysis

[2] Netflix Silverlight Problem. Web. 9 April 2014. http://social.msdn.microsoft.com/Forums/silverlight/en-US/4348d3d1-bc15-41de-8ca6-a06c5e74be94/netflix-silverlight-problems

[3] Google Chrome crash. Web. 9 April 2014. https://support.google.com/chrome/answer/142063?hl=en

[4] Analyzing the Social Web, 1st ed., 2013 by Morgan Kaufmann and Jennifer Golbeck

[5] Big Data Analytics, Sachchidanand Singh, Nirmala Singh, International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India, (2012)

[6] Apache Hadoop. Retrieved on 9th April 2014 from website: http://hadoop.apache.org/

[7] JSON (JavaScript Object Notation). Retrieved on 9th April 2014 from website: http://en.wikipedia.org/wiki/JSON

[8] Apache Hive. Retrieved on 9th April from website: http://hive.apache.org/

[9] Hive SerDe (Serialiser/Deserializer). Retrieved on 9th April from website: https://cwiki.apache.org/confluence/display/Hive/SerDe

[10] Construction of a Sentimental Word Dictionary by Eduard C. Dragut, Clement Yu, Prasad, Weiyi Meng in Conference on Information and Knowledge Management (CIKM) 2010 at Toronto, Canada.

[11] HDFS (Hadoop Distributed File System). Retrieved on 9th April 2014 from website: http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

[12] MapReduce. Retreived on 9th April 2014 from website: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html#Purpose

[13] Apache oozie. Retrieved on 10th April from website: https://oozie.apache.org/

[14] Cloudera Manager. Retrieved on 10th April from website: http://www.cloudera.com/content/cloudera/en/products-and-services/cloudera-enterprise/cloudera-manager.html

[15] Twitter API. Retrieved on 10th April from website: https://dev.twitter.com/docs/api/streaming

[16] Apache Flume. Retrieved on 10th April from website: http://flume.apache.org/

[17] Exploiting Emoticons in Sentiment Analysis by Alexander Hogenboom, Daniella Bal and Flavius Frasincar in ACM, New York, 2013.

[18] Hive-QL (Hive Query Language). Retrieved on 10th April from website: https://cwiki.apache.org/confluence/display/Hive/LanguageManual

[19] The Effect of Negation on Sentiment Analysis and Retrieval Effectiveness by Lifeng Jia, Clement Yu, Weiyi Meng in Conference on Information and Knowledge Management (CIKM) 2009 at Hong Kong.